

Share issues

The US National Institutes of Health is toughening its funding rules to persuade researchers to share materials more widely. The move is commendable but it raises critical questions that urgently require resolution.

Progress in science depends on replication of results, so there is widespread support for the principle that researchers should share the materials needed to reproduce their published claims. Many journals, including *Nature*, specify this as a condition of publication, and several funding agencies — including the National Institutes of Health (NIH), the Howard Hughes Medical Institute and the Wellcome Trust — expect their researchers to make materials widely available. Even so, failure to share is still common. So we should welcome the NIH's announcement of stronger measures to promote 'good citizenship' in the life-science community.

As reported on page 953, the NIH will soon require grant applications to include a specific plan for sharing model organisms and materials that may result from the funded research. Applicants' track records of sharing will also be taken into consideration when their grants are up for renewal. This policy seems to have teeth, but the NIH has so far published only a brief summary statement. The rules will take effect in October, so there is some urgency to clarify exactly how they will work in practice.

Researchers have many motives for refusing to share published materials. Some — such as a desire to maintain a competitive advantage or to use sharing as a way to leverage honorary co-authorship — are contrary to the spirit of publication, and the NIH is right to reject these excuses. But other obstacles to sharing are not so easily dismissed and must still be overcome if the policy is to be effective.

The cost and administrative burden of responding to requests for materials can be substantial, particularly for mice and other

animals that are subject to stringent shipping regulations. The NIH has promised to make supplementary funds available to cover these costs, which should help. But academic labs cannot be expected to run animal breeding facilities, so the NIH must also provide strong support for repositories such as the Mutant Mouse Regional Resource Centers, and ensure that they have sufficient capacity to receive and distribute strains in an efficient and timely manner.

Critics of the NIH policy are understandably concerned about intellectual-property rights, but universities are not disinterested observers in the debate. They are entitled to patent their researchers' potentially valuable inventions, and in some cases the imperative to share may clash with the profit motive. In announcing that it will hold not only individual researchers but also their institutions accountable for sharing, the NIH has staked out a strong position. It remains to be seen how it will be implemented.

Refusals to share resources are almost certainly underreported, because researchers whose requests are denied are often reluctant to make a fuss, especially if they believe that nothing can be done. But unless they are willing to report problems, nothing will change.

Journals have a responsibility to pursue complaints, of course, but the real power lies with funding agencies. The NIH has provided a lead; other agencies should now follow to establish consistent and enforceable community standards. These must be embodied in clear policy statements, which must also include contact details for individuals to whom problems can be referred, and clear explanations of how complaints will be handled. ■

A matter of interest

The biotech industry's top lobbyist faces a problem he was quick to raise as a Congressman: apparent conflict of interest.

Congressman James Greenwood (Republican, Pennsylvania) is, on the face of it, a terrific catch for the Biotechnology Industry Organization (BIO), which last month announced his appointment as its next president (see *Nature* **430**, 495; 2004).

Greenwood was, until the announcement, chair of the investigations subcommittee of the House of Representatives energy and commerce committee. His powerful subcommittee has this year been vigorously investigating alleged conflicts of interest among senior scientists at the National Institutes of Health (NIH).

The main outcomes of his investigation, so far, have been twofold: the implementation of stringent rules blocking links between senior NIH scientists and the biotechnology industry, and the tarnishing of the biomedical agency's hitherto pristine image. Some accounts of the hearings have portrayed the NIH as corrupt, spoiling the chances of any further increases in its admittedly generous budget.

NIH researchers are somewhat flummoxed by news that the man leading the investigation into their links with the biotechnology industry is about to work for it. It looks to them as though Greenwood was sternly demanding that publicly employed scientists steer clear of private-sector involvement, even as he prepared to cash in on his own public service by accepting an annual salary of \$650,000,

plus bonuses, to lobby for an industry that he used to help supervise.

The committee's investigation — sparked by allegations first reported in the *Los Angeles Times* — found that a small number of individuals at the NIH had succumbed to the temptation of the barrage of offers from industry, and accepted consultancy or speaker fees. Most had done so within their employment rules; a few appear to have breached them.

But the hearings of Greenwood's committee created the impression that links between NIH researchers and biotechnology companies were out of control and corrosive to the public interest. This has created a backlash that is likely to constrict such links for years to come. The integrity of government scientists is a serious matter and an appropriate one for congressional investigation. But it is difficult to see Greenwood's career change as being consistent with the moral thrust of his committee's work.

He has resigned as chair of the committee and will leave Congress in January to work for BIO; under House rules, he is not supposed to lobby former congressional colleagues directly for a year after that. His appointment nonetheless leaves the impression that mid-level and senior NIH researchers are being held to far higher ethical standards than those adhered to by their overseers. ■





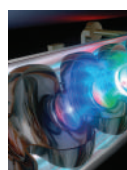
Lost lander

Beagle team sniffs out reasons for the failed mission
p954



Stymied

Warnings about bird flu overlooked in Chinese journals
p955



Cool machine

Physicists lay plans for linear particle collider
p956



Flying start

Ornithologists branch out to track Canada's birds
p957

Universities unnerved by revised rules for sharing NIH research

Charles Jennings

Memo to US researchers: read the small print carefully before filing your next grant application to the National Institutes of Health (NIH).

Unknown to many, the rules are changing on 1 October in a bid to encourage grant holders to share the fruits of their research. But university representatives fear that the new rules lack clarity and will cause confusion and delays for this year's applicants.

Come September, when people finalize their applications for autumn deadlines, "there's going to be a big crunch", says Robert Hardy, spokesman for the Council on Governmental Relations (COGR), which represents 150 US research universities in their dealings with funding agencies.

The NIH already expects grant holders to share materials and data with other researchers after publication, but from October, applicants must give explicit details on how they intend to do this. The policy is aimed at the sharing of model organisms, although it will also cover cell lines, associated reagents and protocols.

Under the new rules, applicants will be able to apply for extra funds to cover the sometimes substantial costs of sharing, such as depositing model organisms in public repositories. But a failure to share could lead to penalties: the NIH says that it will take sharing records into account when making future funding decisions.

Share deal

Eric Campbell of the Institute for Health Policy at Massachusetts General Hospital in Boston applauds the new policy. "The two fundamental underpinnings of science are replication and peer review," he says. "If you hamstring one of these, you are hurting science."

Campbell believes that funding organizations have a responsibility to promote a culture of sharing. Two years ago, he surveyed some 1,200 academic geneticists and found that almost half of them had been



Open access: under new rules, researchers who seek an NIH grant must detail their plans for sharing model organisms.

denied requests for information or materials in the previous three years — as a result, 28% of them were unable to replicate published findings (E. G. Campbell *et al.* *J. Am. Med. Assoc.* **287**, 473–480; 2002).

But the COGR is concerned about the way in which the NIH rules will be interpreted. When the changes were announced — fairly quietly in early May — the council wrote an open letter to the agency listing a number of concerns, including the lack of public consultation and a failure to deal with intellectual-property issues. Council president Katharina Phillips added that the NIH had not clarified what constitutes an acceptable sharing plan or the legal basis for the policy. Rather than insisting on detailed plans in grant applications, the COGR wants the NIH to accept a simple statement of intent to share materials.

The NIH has promised to clarify its policy in a document soon, but this has not yet been published, and the COGR says that the agency should delay its October deadline until the concerns have been resolved. The NIH says that it remains committed to its timetable.

The rule change is in line with a report published last year by the US National Research Council, part of the National Academies, which argued that stronger community standards for sharing were needed. "The NIH are moving in the right direction," says Thomas Cech, president of the Howard Hughes Medical Institute in Chevy Chase, Maryland, and chairman of the committee that produced the sharing report.

Balancing act

Observers, however, point out that there is an inherent tension between the NIH, which seeks to maximize the impact of the research it supports, and the universities, which are entitled to patent and profit from inventions made with government funds. In some cases, such as that of frequently requested reagents, it may be in all parties' interests to license the materials to companies that supply them to other researchers. Under the new policy, NIH reviewers will then have to consider the terms under which materials will be available and determine whether the price is fair.

Licensing decisions are normally the responsibility of universities rather than individual researchers, and the NIH policy statement also warns that institutions will be held accountable for the sharing behaviour of their investigators. The risk of losing NIH funding represents a considerable threat, but proponents of the policy see this as a last resort. "We will use discretion," says Norka Ruiz Bravo, deputy director for extramural research at the agency. "We don't want to use a hammer to swat flies."

► <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-04-042.html>

Analysis highlights suicide risk of antidepressants

Erika Check, Washington

The potential link between antidepressants and suicidal behaviour in children has been given further substance by a study from the US Food and Drug Administration (FDA).

Children and adolescents on antidepressants called selective serotonin reuptake inhibitors are 1.8 times more likely to experience suicidal thoughts or behaviour than those who do not take the drugs, the FDA said on 20 August. The analysis, by Tarek Hammad, an FDA safety reviewer, rests on drug-company data from 25 clinical trials.

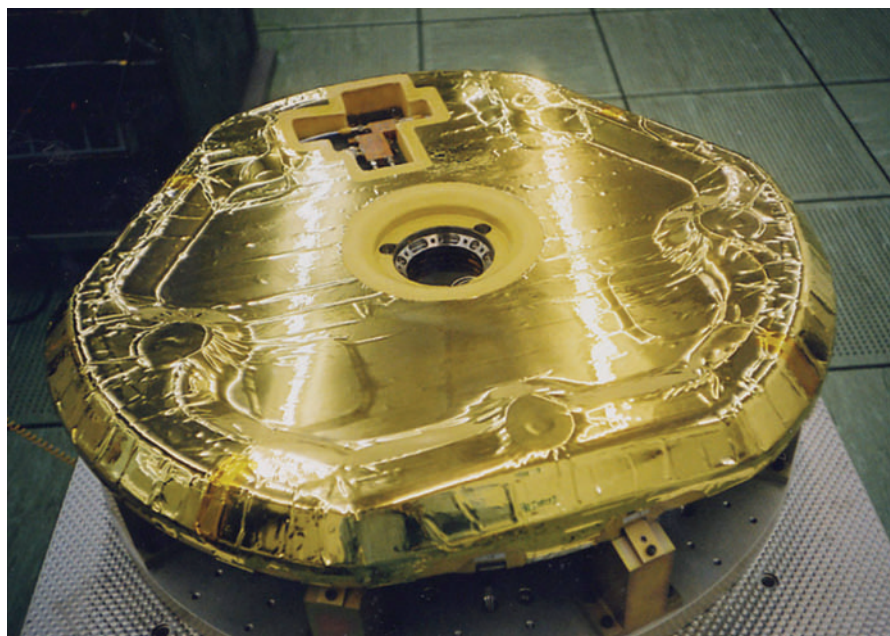
The study confirms earlier findings by Andrew Mosholder, an epidemiologist at the FDA, who produced an analysis of the data in February but was banned from presenting his results in public. The FDA said at the time that the findings were only preliminary, but news of the suppression fuelled fears that data on the drugs were being withheld by drug companies and government regulators.

Mosholder's analysis put the risk only slightly higher, at 1.9 times greater for children on the drugs. But unlike that study, Hammad's analysis did not find a significantly higher risk for the subgroup of children with serious depression.

The risk also varied for the nine drugs studied. Fluoxetine, marketed as Prozac, did not increase the risk of suicide. Others, such as venlafaxine (Effexor), were linked to a significantly higher risk of self-harming thoughts and behaviour.

The analysis increases the pressure on the FDA to further discourage doctors from using the drugs 'off-label' — prescribing them in cases for which they are not approved. Of the nine drugs in the study, only fluoxetine is specifically approved to treat childhood depression. The FDA has already asked makers of the drugs to use labels warning that adults and children on the medication must be monitored for suicidal behaviour, but two advisory committees will meet next month to discuss whether more should be done.

The data used in the study are largely unpublished, leading to renewed calls for the government to require drug makers to release results from clinical trials. "Mosholder's initial findings have been confirmed and reconfirmed," says Senator Chuck Grassley (Republican, Iowa). "When there are known risks with a drug, the information must be made available in a timely way."



Missing in action: the Beagle 2 martian lander was dogged by money problems.

Beagle team hounds space agency over lost lander

Mark Peplow, London

The team behind the ill-fated Beagle 2 Mars lander has hit back at its critics in the European Space Agency (ESA).

The £40-million (US\$72-million) lander was last seen leaving ESA's Mars Express spacecraft for the surface of Mars on 19 December last year, since when little information has emerged about why it went missing. A largely confidential ESA report, issued in May, included a list of criticisms of design choices made by the lander team — but now the Beagle researchers have had their say.

The team's report into what went wrong, released on 24 August, says that ESA regarded Beagle as nothing more than an instrument on board its Mars Express craft. The programme "was consistently forced to meet the needs of Mars Express at the expense of its own requirements", write mission manager Mark Sims of the University of Leicester and his colleagues.

The report also dismisses accusations by some engineers that Beagle's airbags, which were meant to cushion its impact on the martian surface, were not properly tested. The team points out that the airbags were made by the company behind a similar system used on NASA's exploration rovers, which arrived safely on Mars this January.

The Beagle researchers say they could have learned more about what went wrong had they not been prevented from installing a communication device that would have operated after the lander left Mars Express and before it touched down on the planet. They had designed such a system, but the

report says they were advised by ESA that no Earth or space-based receiver would be available to relay the messages.

One possible clue to the mission's failure, says the report, comes from readings from Mars Express, which show that the martian atmosphere was thinner than expected at the point at which Beagle was released. That means Beagle's descent would not have been slowed as much as calculated. Sims and his team say that this problem could not have been foreseen, and add that a better understanding of the martian atmosphere is crucial to the success of any future missions.

Despite the antagonistic tone taken towards ESA, the Beagle team agrees with the agency on several points. Both parties say that any future orbiter-and-lander mission should be managed as a single programme. Sims and his colleagues also acknowledge that Beagle suffered from a lack of money early in the project: just 20 months before the lander was to be attached to Mars Express, 80% of the craft remained unfinished, the report admits.

ESA officials had only just seen the report as *Nature* went to press, but they say that it will be of great use to future Mars lander teams. "I feel glad that they have not homed in on a single failure mode," says David Southwood, director of science at the agency. The fact that no single fault has been focused on "has forced an in-depth analysis of everything".

For more news and analysis go to
news@nature.com
www.nature.com/news

Bird flu data languish in Chinese journals

David Cyranoski, Tokyo

Potentially alarming findings on the avian influenza epidemic currently sweeping southeast Asia went largely unnoticed because they were published in Chinese-language journals, it emerged last week.

Chinese researchers reported this January that the strain of the bird flu virus involved, known as H5N1, had infected pigs. This is a cause for concern, as some virologists believe that previous human flu pandemics, such as the one in 1918 that killed millions, originated in pigs. Yet neither the World Health Organization (WHO) nor the UN Food and Agriculture Organization (FAO) was aware of the results when they were published. Last week both groups were hurriedly trying to get the papers translated.

The findings became widely known only when an author on the papers, virologist Chen Hualan of Harbin Veterinary Research Institute in northeast China, presented them at the International Symposium on the Prevention and Control of SARS and Avian Influenza in Beijing on 20 August. The results, which show that the virus was found in 2003 in pigs in southeast China, were originally published in January (L. I. Haiyan *et al. Chin. J. Prev. Vet. Med.* 26, 1–6; 2004). A second paper by Chen and her colleagues appeared in May, reporting similar results



Rotten swine: pigs could host viruses, brewing a human pandemic.

from pigs tested elsewhere in 2001 and 2003 (L. I. Haiyan *et al. Chin. J. Vet. Sci.* 24, 304–309; 2004).

The H5N1 strain, which can be fatal to humans, does not jump easily from birds to humans or between humans. But this might change if pigs are infected, says Malik Peiris, a virologist at the University of Hong Kong. The virus could exchange genes with the human flu strains that infect pigs and combine to become far more infectious to humans, he warns. Peiris's group has already shown that a human virus, H3N2, can coexist with other avian influenza viruses and is

widespread in pigs from southeast China. Pigs are an "ideal mixing vessel", he says. "It is certainly sufficient to make me concerned."

The H5N1 virus has spread through most of southeast Asia this year, resulting in the destruction of over 100 million chickens, including 9 million in China. In Vietnam and Thailand, 37 human infections have been reported, with 26 deaths.

The abstract to Chen's January article ends with the warning that "urgent attention should be paid to the pandemic preparedness of these two subtypes of influenza". But the results attracted no attention outside China before last week. The WHO and the FAO are now waiting to see a translation of the article and to hear

from the Chinese agriculture ministry before deciding what steps to take. Until then, they urge caution and say that the significance of the findings cannot yet be judged.

The ministry, which seemed to be surprised by the disclosure, moved to calm fears after the news broke. A ministry press release, issued on 23 August, stated that tests on "over 1.1 million samples of poultry and some samples of pigs" conducted this year showed no evidence of H5N1. Officials add that routine testing by Hong Kong researchers of imported pigs had also not turned up evidence of H5N1.

Greenland ice sheet to get underhand inspection

Amanda Haag

Marine scientists are to take a second stab at a world first — the exploration of the dark waters below the permanent polar ice sheets.

Early next month, the robot submarine Autosub will slide beneath the fringes of the ice sheet that blankets Greenland. Researchers hope to use the vessel to study the underside of the permanent ice, to investigate rising sea levels.

The mission will be the vessel's second such attempt; Autosub was to be deployed under Antarctica's Pine Island Glacier in early 2003, but the ship on which it was being transported was blocked by rapidly drifting ice. Robot subs have previously journeyed under sea ice, which breaks up and reforms annually, but none has attempted to go under permanent ice.

Sea ice is often only a few metres thick, but the ice shelf that Autosub will explore is 200 metres deep and likely to contain narrow channels in which the sub could get irretrievably stuck or lost. In addition,



Robo drop: Autosub will be released next month to study the Arctic's ice from below.

unexpected changes in salinity could create buoyancy problems that might sink the vessel or force it up into the ice. Strong currents could also send the sub off course.

"This very much is undersea space exploration," says Ken Collins of the Southampton Oceanography Centre and the expedition's scientific coordinator. "If anything goes wrong far out there, we won't see it again." Collins and his colleagues on

board the *James Clark Ross* have spent the past two weeks sending Autosub on only slightly less perilous missions under the transient sea ice, awaiting the return of the vessel — worth roughly £1.5 million (US\$2.7 million) — like worried parents.

The 7-metre-long robot, powered by some 5,000 household batteries, is programmed to map the contours and biology of the sea bed, to measure the thickness of ice overhead and to collect sea water for analysis back on the ship. Together, the results should improve understanding of how fresh water from the ice is cycled into the ocean, contributing to rises in sea level.

"The effects of global warming are greatest and fastest in the Arctic," says Peter Wadhams, a polar scientist from the University of Cambridge, UK, who is leading part of the expedition. "The Arctic ice has decreased in thickness by 40% over the past 20 years, so by 2080 the Arctic will be ice-free in summer."

www.soc.soton.ac.uk/au

Brickbats for fossil hunter who claims skull has false tooth

Rex Dalton

Almost every important fossil hominid seems to spawn controversy. Now Toumaï, man's oldest known ancestor, has his.

The integrity of the 7-million-year-old specimen (*Sahelanthropus tchadensis*) came under attack in April, when the *South African Journal of Science* published an article claiming that Toumaï's jaw had a molar glued in the wrong place to round out his set of teeth (A. Beauvilain and Y. Le Guellec *S. Afr. J. Sci.* 100, 142–145; 2004).

The article was written by Alain Beauvilain, a geographer from the University of Paris X and an author on the original paper, with Yves Le Guellec, a dentist acquaintance of his. Palaeo-anthropologists say there is nothing wrong with the published description of Toumaï (M. Brunet *et al. Nature* 418, 145; 2002), which puts pressure on the South African journal to clarify matters.

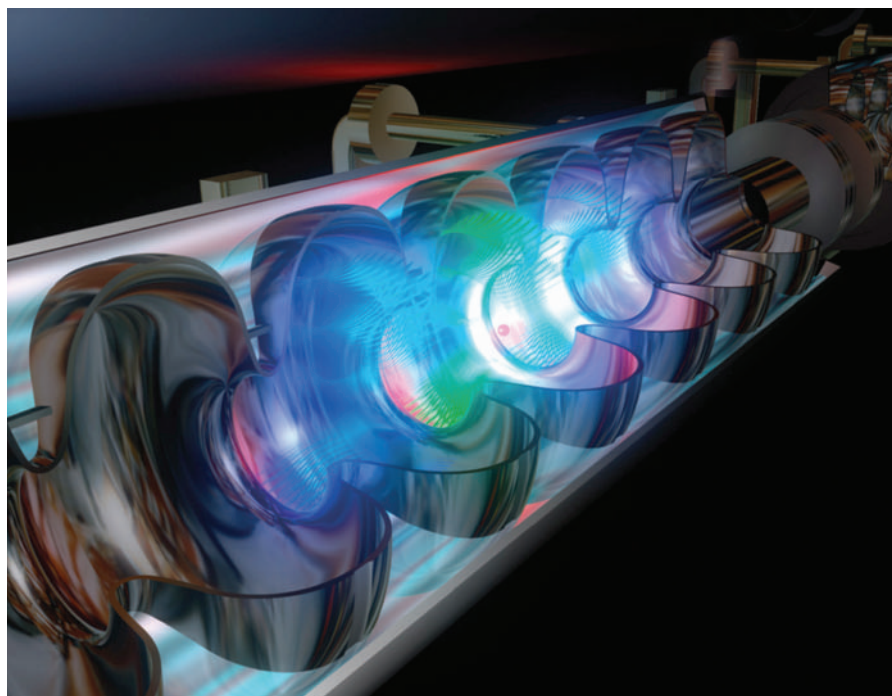
The backlash against the pair, who attracted much attention in the French media when their article was published, began in June. Martin Pickford, chairman of palaeoanthropology and prehistory at the Collège de France in Paris and the reviewer of the article, wrote to the *South African Journal of Science* offering his "humble apologies" to the journal and to the leader of the Toumaï discovery team, Michel Brunet, a palaeoanthropologist at the University of Poitiers. Pickford admits that he "paid insufficient attention" to details in the article.

Protests continued last month, when more than two dozen of the world's leading palaeoanthropologists wrote to the journal saying that the third molar on the right of Toumaï's jaw is indeed a right molar — not a left one as the critics claim. Brunet says his analysis shows that Toumaï's right molar matches breaks in the tooth's roots, apparently settling the argument.

Beauvilain has helped Brunet organize many trips to bandit-infested areas of the Sahara, but he acknowledges that he was not involved in writing the manuscript and is not a trained palaeoanthropologist.

The South African journal is trying to resolve the dispute, although it has not been able to contact Le Guellec, and Beauvilain is sticking to his guns. "The tooth is clearly from the left side, even if a number of palaeoanthropologists maintain it is from the right," he told *Nature* last week. ■

Additional reporting by Sabine Louët.



Along the right lines: Germany's niobium accelerator will lead research on the Higgs boson.

Superconductor beats copper in plans for particle collider

Geoff Brumfiel and David Cyranoski

Physicists have chosen the technology that will be used in the next generation of particle collider, an international experiment that hopes to study the particle thought to endow all others with mass.

The collider will use superconducting technology rather than copper coils, the International Committee for Future Accelerators decided on 20 August. The move, which comes after years of debate, means that hundreds of scientists in the United States and Japan, who have been working on the more traditional copper devices, must refocus their efforts.

Support for an international linear collider has been building over the past four years. The machine would smash together electrons and positrons, their antimatter equivalent, at enormous energies. This could generate the elusive Higgs boson, which theory predicts interacts with other subatomic particles to give them mass.

"You can think of the linear collider as a precision microscope," says Philip Burrows, a physicist at Queen Mary, University of London. "It is a tool that will allow us to really pin down the properties of the Higgs."

Two technologies have been front runners from the outset. Researchers at DESY, the German national research centre for particle physics in Hamburg, have promoted a design that uses superconducting niobium to create the huge electric fields for accelerating electrons and positrons (see *Nature* 410, 397;

2001). A rival system, which uses copper coils, was developed at KEK, the high-energy accelerator research organization in Tsukuba, Japan, and at the Stanford Linear Accelerator Center (SLAC) in California.

"This was a really difficult decision," says Barry Barish, head of the technical team charged with reviewing the two proposals. "Both technologies had their pluses and minuses." Copper coils, for example, are capable of accelerating the particles over a shorter distance, but the superconducting technology would use less energy. In the end, Barish says, the team settled on the superconducting technology for a variety of different reasons, including its ability to create beams with a higher density of particles, resulting in an increased collision rate.

The move is likely to help DESY, which has struggled in recent months to get funding from the German government for a smaller prototype machine. But it will also force researchers at SLAC and KEK, which until now have devoted millions of dollars to developing copper technologies, to change direction quickly. Still, the director of SLAC, Jonathan Dorfman, says: "Much of the progress will be able to be redirected."

With the technology chosen, Barish says, the international community will start designing the machine, and hope to deliver a preliminary blueprint by 2007. Although the cost of the machine is not yet known, Barish estimates that it will be "billions of dollars". ■

Director's salary makes chemists see red

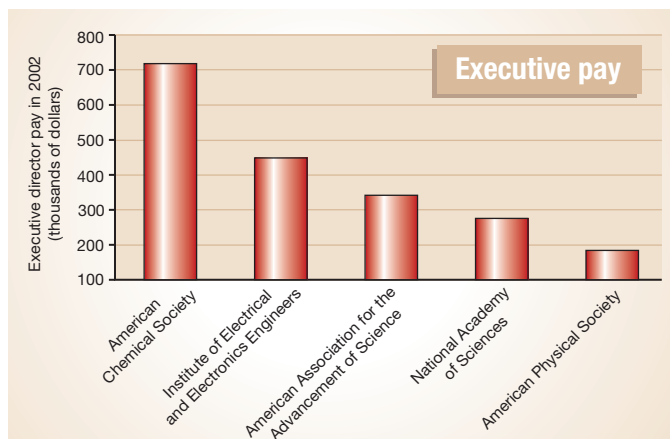
Geoff Brumfiel, Washington

The size of the pay package awarded to a former executive director of the American Chemical Society (ACS), which was nearly \$768,000 in 2002, has led to protests from members.

According to income-tax papers filed by the society, John Crum received compensation of around \$721,000 in 2002, the penultimate year of his 20-year tenure. Crum was also rewarded by roughly \$33,000 in contributions to benefit plans and a \$14,000 travel expense account, taking his pay far above the directors of other scientific societies such as the National Academy of Sciences and the American Association for the Advancement of Science (see chart).

Ten chemists, including Nobel laureate Roald Hoffmann of Cornell University in Ithaca, New York, published a protest letter in the 9 August issue of *Chemical and Engineering News*, the weekly magazine of the ACS. "It is difficult to believe that ACS could not find a capable executive director for less than three-quarters of a million dollars a year," they wrote.

"To many of our members some of the pay of our executives really looks off the



scale, and I understand that people can feel that way," says James Burke, who chairs the society's board of directors, which approves executive salaries. But he says that pay at the ACS, considered to be the world's largest scientific group, is in line with other organizations of similar size.

"The American Chemical Society is an integrated and complex organization," says Burke. "We have to make sure that we are competitive with private publishers, such as Elsevier and Wiley." He adds that the salaries of ACS executives are analysed every two years by outside consultants and that they are consistently found to be in the appropriate

range for such an organization. Commercial publishers do not disclose salary details, but executive directors are thought to earn in the region of \$500,000 to \$1.5 million.

Robert Bergman, a chemist at the University of California, Berkeley, and the lead author of the complaint letter, disagrees with Burke. A consultant being paid to determine the salary of a client might be inclined to choose the highest salary possible, he says. "I feel that this whole arrangement has a built-in conflict of interest," he adds.

Daniel Borochoff, president of the American Institute of Philanthropy, a non-profit charity watchdog in Chicago, Illinois, says that the pay may seem unusually high, but it is acceptable if the ACS can prove that the salary is competitive. "The membership may be in an uproar," he says, "but the Internal Revenue Service considers a salary reasonable if it is comparable to what gets paid in either a non-profit or a for-profit job."

Burke expects the issue to be raised at the ACS's annual meeting, held this week in Philadelphia. "I would expect that some members will want some more clarification," he says. "And we'll do our best to provide that." ■

Canada rings the changes for study of northern birds

Erika Check

Ornithologists are laying high-tech plans to study the feathered forest dwellers of Canada's northern wilderness.

Billions of birds live in the country's 6 million square kilometres of boreal habitat. The rugged isolation protects around 200 species, but also stymies biologists who want to study them. And the need to understand the area's bird life has become urgent, as oil, gas, timber and mining firms begin to stake claims in the forests.

The Canadian Wildlife Service plans to use innovative techniques, many new to ornithologists, to research the region's birds. "In the past few years, there have been tremendous advances that suddenly make a study of this scale and scope a realistic possibility," says Mike Norton, a boreal biologist with the Canadian Wildlife Service in Edmonton. He discussed the plan at a joint meeting of the American Ornithologists' Union and the Society of Canadian Ornithologists in Quebec City, Canada, on 16–21 August.

The use of one new technique, stable



On song: the white-throated sparrow is a familiar summer resident of northern forests.

isotope analysis, has been pioneered by Norton's ornithologist colleague Keith Hobson, based in Saskatoon. After trapping a bird, he plucks a single feather and measures the amounts of the different isotopes of hydrogen in the sample. As the levels of individual isotopes differ in water from different parts of the forest, "this lets us define where the birds we capture are coming from, and also where they spend the winter at the end of their migration," says Hobson.

Other new tools include improved

omnidirectional microphones to collect faithful sound recordings in the wilderness. Such equipment can help make up for a dearth of trained ornithologists able to go to the boreal forest to identify songbirds, says Norton. And, he adds, better ways of modelling statistical data will help researchers interpret the information they collect.

Norton estimates that the scheme will cost about Can\$15 million (US\$11.5 million) a year and will take a decade to plan fully, although he hopes to start implementing parts of it over the next few years.

Researchers say that the new techniques must be ramped up before much more of the wilderness disappears. They point out that the region is crucial for birds across the entire North American continent. In May 2003, the non-profit group Bird Studies Canada, based in Port Rowan, Ontario, found that one out of every three North American land birds breeds in the boreal region. But the same study also raised a red flag — it reported that 40 boreal species are in decline. ■

Monkey shortage causes delays for medical researchers

Washington Life-saving research is being slowed by a shortage of lab primates, according to a global audit in the current issue of the *American Journal of Primatology* (63, 225–237; 2004).

Hans-Erik Carlsson from Uppsala University in Sweden and his colleagues looked at nearly 3,000 papers published in 2001, which together used more than 41,000 animals. But the team reports that many experiments require a specific type of primate, which can often be in short supply. This can delay research and means that some animals are repeatedly used for different experiments. This in turn can affect the outcome of some studies, say the authors.



Primate number: a lack of monkeys means that some will be used in several experiments.

More disturbingly, 86% of the studies did not disclose the history of their primate subjects, making it difficult to repeat the experiments or assess the results. “The re-use of animals is a danger in any experiment,” says William Morton, director of the Washington National Primate Research Center in Seattle. “I think that, more and more, the scientific literature will include the characteristics and history of the particular animal.”

Carlsson says that journal editors may delete information about primate histories even when authors include it, considering it extraneous.

Anthropologist turns heads with mystery dates

Munich A German anthropologist is facing accusations of misconduct from the University of Frankfurt, after concerns came to light about his research and apparent attempts to sell university property.

According to the German news magazine *Der Spiegel*, Reiner Protsch von Zieten carbon-dated several human skulls from Germany and found them to be up to 30,000 years old. Other labs date the bones at just 7,000–8,000 years old. Protsch von Zieten’s

Cash injection helps China line up HIV drugs

Tokyo As queues grow in China for access to drugs to fight HIV (see right), the country’s clinics received a US\$32-million boost on 19 August with a grant from the Global Fund to Fight AIDS, Tuberculosis and Malaria.

The money will be administered over two years by the China Comprehensive AIDS Response (China CARES), a community-based network of clinics created by the government last year. The cash should initiate a five-year US\$165-million programme, co-funded by China and the Global Fund, to combat HIV.

The project aims to make antiretroviral medicines more readily available. Of an estimated



840,000 HIV patients in China, only about 3,000 currently have access to the drugs. The project aims to raise this to 40,000 over five years. The government has also licensed two Chinese companies to make generic antiretrovirals, which should lower the cost of the drugs.

results have been used to back controversial hypotheses on prehistoric population movements in central Europe. It is unclear how Protsch von Zieten managed to get such radically different results; the university says that he stands by his findings.

Protsch von Zieten was suspended from the university in April, after allegedly attempting to sell more than 200 institute-owned ape skulls to collectors in the United States. According to the university, Protsch von Zieten says he owns the skulls himself. Legal proceedings in this case are pending.

Nature was unable to reach Protsch von Zieten for comment.

Parents upset by mass removal of child organs

London A long-running furor regarding the removal of organs from deceased children without parental consent has heated up in Ireland.

In the past few weeks, several Irish hospitals have revealed that they supplied pituitary glands to pharmaceutical companies in the 1980s. Some have confirmed that, in accordance with their practices at the time, consent was not sought for this. The companies often paid a small fee for the organs and used them as a source of human growth hormone — a product that is now made synthetically.

Novo Nordisk, a Danish-based company, said last week that it received some 7,500 pituitary glands from 32 Irish hospitals in the 1980s. Protest groups say this adds to the need for further investigation.

The Irish group Parents for Justice, which represents parents of the deceased children, claims to have evidence that these revelations are just the tip of the iceberg. Some of the deceased children “had almost every organ in their body removed”, claims the group’s chairwoman, Fionnuala O’Reilly.

An inquiry into the matter, set up by the government in 2002, is expected to report by the end of the year.

Biotech firms settle up over HapMap contract spat

San Diego Two biotechnology firms last week settled litigation that was casting a cloud over the International HapMap Project, a US\$120-million effort to identify how human genetic variation affects disease.

San Diego-based Illumina was accused last year of not fulfilling contractual obligations to Connecticut-based biotech giant Applera (see *Nature* 423, 470; 2003). The two companies were involved in a 1999 joint venture to develop a genetic analysis system. But after the agreement went sour, both developed competing technologies. Illumina’s system is now used for at least half of the HapMap project. Illumina has agreed to return \$8.5 million of \$10 million that Applera’s subsidiary company Applied Biosystems provided for the joint venture.

Correction

The 12 August 2004 News article ‘Victims hit out at university over handling of harassment cases’ (*Nature* 430, 711; 2004) included the following errors.

Officials at the University of Toronto did not ask Lorilyn Benoit inappropriate questions after she made accusations of sexual harassment. In addition, Benoit did not bring her harassment claim against the university. Nor did the university ask Benoit to sign a confidentiality clause. In both cases, Benoit alleges that the officials involved worked at a teaching hospital affiliated to the university but not under its control.

The investigation into Gwen Schwartz’s allegations was discontinued when she and the other parties started to mediate the settlement that they later reached. The University of Toronto has asked us to make clear that the mediator’s report stated: “although the facts remained significantly in dispute, the mediation was successful and all parties were able to reach a comprehensive, final and binding settlement, fully resolving all issues amongst them”.

Nature also wishes to state that Claude Cantin is employed by the teaching hospital, not the university.

The accidental spy

Hiroaki Serizawa's promising US academic career was ruined when a favour to a friend led to him being charged with economic espionage on behalf of Japan. He tells his story to David Cyranoski.

In July 1999, Hiroaki Serizawa, then a cancer researcher at the University of Kansas, made a fateful decision: he agreed to let a long-time friend, Alzheimer's disease specialist Takashi Okamoto, store some research materials in his lab freezer. At the time, about a week after the birth of his first child, Serizawa says he gave little thought to the request. But his decision culminated in his conviction in a federal court, and the end of his academic career.

The episode saw the first application of a clause from a 1996 US law drafted to target economic spies working for foreign organizations. Legal experts suggest that the act may need to be changed in the light of federal prosecutors' failure to convict Okamoto, their main target. But Serizawa's experience is a cautionary tale for academic researchers unaware of the extent to which business and legal interests have penetrated their world.

In Japan, where the affair has been front-page news, Serizawa's case has struck fear into the hearts of researchers considering a move to the United States. And the repercussions continue: Serizawa, who wants compensation for his losses, is now squaring off with Okamoto in a Tokyo court.

These legal fireworks were ignited by research on caveolins, a family of proteins active in cell membranes. Okamoto's group at the Cleveland Clinic Foundation in Ohio had found that one of these proteins, caveolin-3, promotes the formation of the plaques that are thought to destroy the brains of Alzheimer's patients (K. Nishiyama *et al.* *J. Neurosci.* **19**, 6538–6548; 1999).

In early 1999, with new funding and a growing library of cell lines and DNA samples to work with, Okamoto was set to pursue an enzyme — yet to be identified — known as α -secretase. This enzyme cuts the plaque-forming protein in half, and so can prevent plaque build-up in the brain.

But in July 1999, Okamoto suddenly ransacked his own lab. He maintains that he did



On trial: Hiroaki Serizawa ended up with a fine and community service — and lost his funding.

this to prevent one of his research fellows from persisting with “irreproducible” and “reckless” experiments that would bring shame to the research group.

Incriminating evidence

The clinic authorities asked the police to investigate, and it wasn't long before the FBI got involved. Its agents did not accept Okamoto's explanation. An affidavit dated 27 August 1999 accused Okamoto of taking or destroying some 1,000 DNA samples and 250 cell lines developed in his group's research. According to the affidavit, Okamoto “took deliberate and calculated steps” to deprive the clinic of valuable materials.

Eventually, in May 2001, Okamoto was indicted on four counts, the most serious being a violation of Clause 1831 of the 1996 Economic Espionage Act, which covers activities to benefit a foreign organization. “The original intent was to target foreign companies, especially those with state sponsorship,

that steal trade secrets,” says Peter Toren, a New York-based intellectual-property lawyer who helped to draft the act. The clause was invoked in Okamoto's case because he was planning to take up a position in Japan with RIKEN, the Institute of Physical and Chemical Sciences.

RIKEN itself was never accused of complicity. But Serizawa, thanks to the contents of his freezer, was charged with conspiracy to commit economic espionage. Sitting on a couch in the lobby of the Tokyo apartment building of a supporter who has helped bankroll his legal team, the quietly spoken biologist told *Nature* how events unfolded.

Confusion and fear

Serizawa and Okamoto met in the late 1980s at the University of Tokyo, and again when each came to do postdoctoral work in Cambridge, Massachusetts — Okamoto at Harvard University and Serizawa at the Massachusetts Institute of Technology. “We shared information on how to write grant applications and he even gave my family medical advice,” says Serizawa.

Serizawa says he didn't think to question Okamoto's motives when his friend asked him to store the materials. “I could not understand that this was a serious problem,” says Serizawa, who was shocked when federal agents showed up at his laboratory on 2 September 1999 and grilled him for seven hours.

In the end, following a plea bargain in which he agreed to testify against Okamoto, Serizawa was charged only with giving false testimony to the FBI in that first interview. The offending statements — which Serizawa corrected later during the same meeting — were an underestimate of the number of vials that he was holding for Okamoto, a denial that he knew his friend had accepted a job offer with RIKEN, and a denial that he had met Okamoto since receiving the materials.

Serizawa blames the mis-statements on confusion and fear. “They showed up with a subpoena, and I had no idea what that was because in Japan we have no such system,” he says. “I was surprised when I realized that I was really under investigation.”

On 28 May 2003, Serizawa was sentenced to a \$500 fine, 150 hours of community service and three years on probation. “I had no idea what an impact this would have on my career,” he says. The American Cancer Society terminated Serizawa's grant; felons are not eligible for funding from the National Institutes of Health within three years of their conviction; and the University of Kansas, where he had worked since January 1997, would not let him do any more research.

Serizawa eventually found employment



AFP/DAVID MAXWELL

Vial plot? Takashi Okamoto (below) was accused of stealing samples from his lab at the Cleveland Clinic (left) after his lab was ransacked.



outside academia. In November 2003, he was appointed chief scientific officer at Life-Science Catalyst Partners, a consulting firm in Palo Alto, California, that aims to build collaborations in biotechnology between Japanese and US companies. Serizawa appreciates the irony of someone who was once accused of being a spy taking such a job. But he cannot forget his ordeal. "I miss research and academia," he says.

Serizawa blames Okamoto for losses including some \$500,000 in legal fees, about \$200,000 of which have been covered by his support group. He is suing his former friend for \$770,000 in compensation. The court will have to sort out contradictory views on what happened. For example, Serizawa says that Okamoto should have stated publicly that his friend had no prior knowledge of the contents of the vials. "His testimony was crucial information," says Serizawa. "No one else knew what happened."

According to Serizawa, his wife called Okamoto and asked for help, but Okamoto hung up on her. Serizawa's support group has the registered mail documentation for a

three-page letter sent — when they couldn't locate Okamoto — to Okamoto's lawyer, ex-wife and parents pleading for him to "appear in public and discuss the truth of the incident". But Okamoto denies that he was ever approached to do anything on his friend's behalf. "I was never asked to come to Serizawa's defence," he asserts.

While Serizawa's legal team was preparing his defence, Okamoto was back in Japan, fighting efforts to have him extradited. Finally, in March this year, a Tokyo court judged that Okamoto's deeds were not carried out to benefit RIKEN — and, for the first time, Japan refused a US extradition request.

Broken friendship

Okamoto had by then been forced to resign from RIKEN. In September 2001, he moved to the Sea of Okhotsk Hospital, in the small town of Tanno at the northern tip of Japan. He still maintains his innocence, and is angry that he spent 57 days in jail awaiting a decision on his appeal against extradition. "I couldn't see my patients, I was deprived of my rights, and I suffered terrible emotional pain," he told *Nature*.

Those complaints cut little ice with Serizawa's supporters. "As a scientist and as a human, Okamoto has the responsibility to apologize to Serizawa," says Ken-ichi Arai, former director of the University of Tokyo's Institute of Medical Sciences.

A decision in Serizawa's suit against Okamoto could come as early as next month. Meanwhile, legal experts are debating what

the case means for clause 1831 of the Economic Espionage Act. Some believe that the difficulty of proving that someone's actions were intended to benefit a foreign organization makes the clause toothless. "It would be hard to convict anyone," suggests Robert Kneller, an expert on intellectual property at the University of Tokyo.

But researchers should not rest easy. Clause 1832 of the same act lists most of the same crimes without the stipulation of benefiting a foreign entity. It carries a lighter maximum penalty, but is being used to convict foreigners.

The 2003 report from the National Counterintelligence Executive underlines the US government's tough stance: "Foreign businessmen, scientists, academics, and government officials from more than 90 countries continued targeting sensitive US technologies and corporate trade secrets in both 2002 and 2003," it noted. The report blamed this in part on the United States' "commitment to global information sharing through academic and scientific exchange".

This could mean closer scrutiny of academic laboratories. "My sense is that, despite the publicity given to several cases, the bulk of the scientific community remains unaware of the Economic Espionage Act," says Mark Frankel, director of the Program on Scientific Freedom, Responsibility and Law at the American Association for the Advancement of Science in Washington DC. Serizawa's advice on that point is clear: ignorance is not bliss. ■

David Cyranoski is *Nature's* Asian Pacific correspondent.



While you were sleeping

Ever woken up with the answer to a problem that had seemed insoluble the night before, or able to perform a task that had previously taxed your skills? We may soon know why, says Laura Nelson.

On 17 February 1869, Dmitri Mendeleev nodded off over a game of solitaire. On waking, he found he had solved the conundrum of how the chemical elements could be grouped in a meaningful way. The periodic table, used by every chemist today, was born.

Legend has it that inspiration came to Mendeleev in a dream, although he probably embellished the story for dramatic effect. But most of us will have experienced similar, if less epoch-making, insights following sleep. And our ability to perform complex tasks, such as playing a musical instrument, can be better the morning after a practice session the night before.

Cognitive scientists group phenomena such as learning to solve logical problems, or improving musical skills, under the heading of 'procedural' learning. This means learning how to do something, rather than simply accumulating a list of facts. For years, researchers have suspected that one important function of sleep is to strengthen this

process. But only now are scientists starting to reveal the specific patterns of brain activity involved. "Sleep is to do with cells and building their connections," says Giulio Tononi, a psychiatrist at the University of Wisconsin in Madison.

In part, the slow progress so far reflects sleep researchers' long-standing focus on REM, or rapid eye movement sleep (see 'The mysteries of the night', overleaf). This coincides with dreaming, during which our brains show patterns of activity that look very similar to those when we are awake. In addition to the trademark flickering eye-balls, both men and women show signs of sexual arousal during REM.

It is hardly surprising, then, that researchers have been drawn to this all-action phase, rather than to the periods of deeper sleep that surround each bout of REM. This quieter phase is known as slow-wave sleep — after the synchronized waves of electrical activity that can be seen on electroencephalograph (EEG) traces undu-

lating across the brain at fewer than four cycles per second.

But despite years of research, the evidence linking REM sleep to procedural learning is contradictory and confusing¹. For example, across species there is no firm correlation

"During training, the brain tags a task as something to be improved on later during sleep."



between the proportion of sleep spent in REM and learning ability. What's more, people who are taking antidepressants that decrease the amount of REM sleep do not have problems with procedural memory. And rare patients with forms of brain damage that eliminate REM completely, similarly show normal procedural learning.

So in the past few years, researchers interested in procedural learning have started to focus on slow-wave sleep — and in doing so have begun to make progress. Perhaps the most convincing evidence linking the quieter phase of sleep to procedural learning comes from Tononi. His team recently published a paper in *Nature*² that for the first time linked local changes in short-wave activity to learning a specific task.

Ride the slow wave

Tononi's human volunteers had 256 electrodes placed over their scalps to take EEG recordings of the electrical activity of brain-cell networks before, during and after sleep. Immediately before the subjects dozed off, they played a computer game requiring them to move a cursor to a position on the screen, believed by the participants to be fixed. But during some sessions, the position was moved for some players, who subconsciously learned how to adapt to this handicap.

Immediately before they dropped off to sleep, the subjects who had learned the compensation task showed heightened slow-wave activity in specific brain areas known to be involved in processing spatial information. This was presumably because of the formation and subsequent strengthening of new connections between nerve cells.

During slow-wave sleep, this activity continued, gradually dying down. Tononi suspects that this decline in slow-wave activity is important in strengthening procedural

"Different stages of slow-wave sleep might be more important for different types of learning, such as how to play a piano or drive a car."



The morning after: sleep seems to yield the biggest benefits on the toughest computer tests.

learning — perhaps because the waking brain is 'noisy', with other signals interfering with the learning process.

After a night's sleep, the volunteers who had learned to compensate for the shifting target were better at the task than they had been the night before. And there was a positive correlation across the subjects between the magnitude of the increase in slow-wave activity and the extent of the sleep-induced gain in performance.

This correlation has excited other researchers in the field³. "This is the first real evidence that the oscillations are for memory. It is very impressive to see the slow waves change locally," says Jan Born, a neuroendocrinologist at the University of Lübeck in Germany. He and others are now looking afresh at their

own experiments on learning, wondering whether it will be possible to tie sleep-induced enhancements to changes in activity in specific parts of the brain.

The past couple of years have brought a rich harvest of findings on sleep and learning that should now be amenable to investigation using EEG and other techniques. Born, for instance, has worked on insights similar to Mendeleev's inspiration about ordering the chemical elements. He asked volunteers to generate a succession of seven sequences of numbers using simple rules. But there was a hidden rule — the second sequence always matched the final one. If the volunteers were allowed sleep between two sessions on the task, they noticed the short cut more quickly⁴. It seems that their sleeping brains processed information that had already been learned. "There has to be a memory that sleep builds upon," says Born.

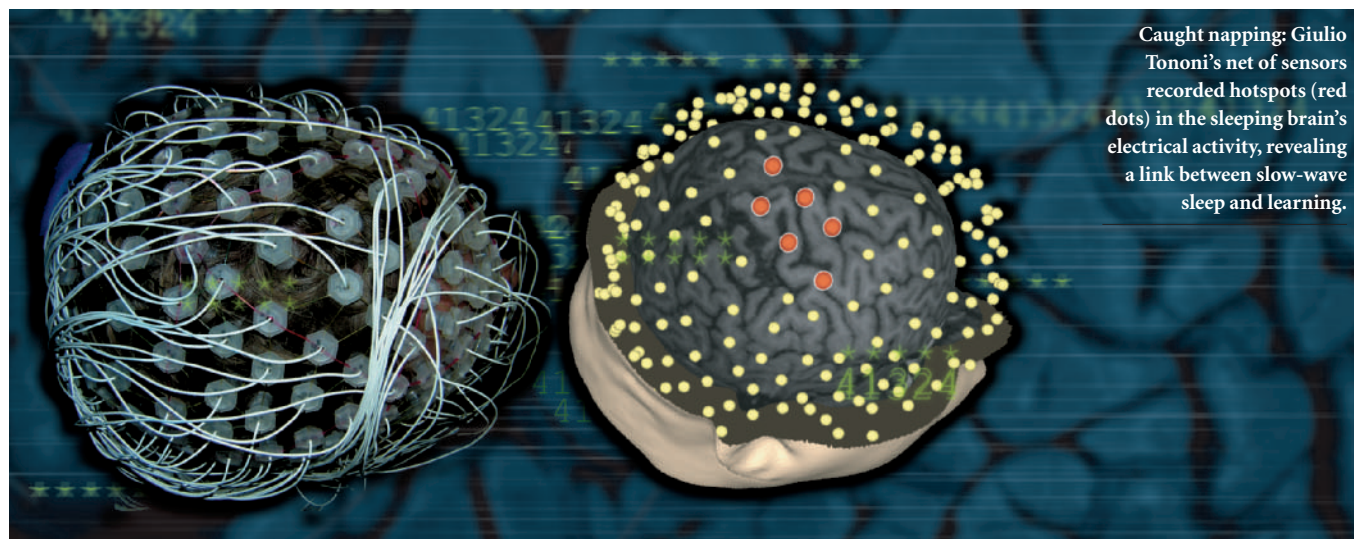
Tag team

Robert Stickgold, a psychiatrist at Harvard Medical School in Boston, has perhaps the richest series of findings awaiting physiological investigation. He believes that, during training, the brain tags the task as something that will be improved upon later during sleep.

Even more impressively, Stickgold's unpublished data indicate that the brain tags tasks in 'chunks', giving higher priority to those that were the most difficult to learn. When Stickgold asked volunteers to type different number sequences on a keyboard, the improvement with sleep was greatest for the sequences that had been the most difficult before sleep. "The brain grabs weaker connections and strengthens those," argues Stickgold.

Such tagging may also help in integrating tasks such as putting together the musical phrases in a piano sonata. "You learn things in pieces and then sleep smoothes them out," Stickgold suggests.

But it seems that memories strengthened



Caught napping: Giulio Tononi's net of sensors recorded hotspots (red dots) in the sleeping brain's electrical activity, revealing a link between slow-wave sleep and learning.

The mysteries of the night

We spend a third of our lives sleeping. But why? Despite recent progress in understanding the role of sleep in strengthening learning, sleep scientists are embarrassed to admit that our need for so much sleep remains largely an enigma.

"Up to now, the science of sleep has not been conclusive," says Sara Mednick, a psychiatrist at the Salk Institute for Biological Studies in La Jolla, California. For a long time, scientists merely speculated about the functions of sleep. It could be required for general body maintenance. Some suggested that the sleeping brain is conserving energy, or recharging its energy stores. Others thought that the brain might be purging itself of damaging chemicals.

There was also an idea that animals sleep to avoid predators. The opposite has since been

shown to be the case, as animals need to sleep so much that they'll do it despite the danger of predation. Some animals, such as ducks, sleep in groups and take it in turns to keep one half of their brains awake while the other half is asleep¹⁰.

The modern era of sleep research can be traced back to 1953, when Eugene Aserinsky, a PhD student at the University of Chicago, spent hours watching the twitching eyelids of sleeping volunteers. He characterized rapid-eye movement sleep (REM), and convincingly linked the state to dreaming¹¹. Scientists were intrigued, and assumed that REM's dynamic nature meant that something useful must be going on. It seemed as if the study of dreams was about to be wrested from Freudian psychoanalysts, and placed squarely in the domain of mainstream science.

Fifty years later, and we're still not much closer to knowing what that the function of REM sleep might be. Despite the contradictory evidence linking REM to learning¹, some cognitive scientists still think that it might be involved in strengthening memories. Others speculate that REM dreams allow us to process information without the restrictive influence of 'rational' thought.

In the past few years, neuroscientists have begun using sophisticated imaging techniques to investigate the brain regions that are activated as we dream. But so far, their work has not produced a convincing theory of dreaming. "There's not much science," concedes Mednick. It will be a while before we know dreams are made of — and what, if anything, they are for.

during sleep become labile for a period after waking: if volunteers learn to tap their fingers in one sequence, and are the next morning asked to learn a second sequence, they seem to lose the sleep-induced boost in performance on the first⁵.

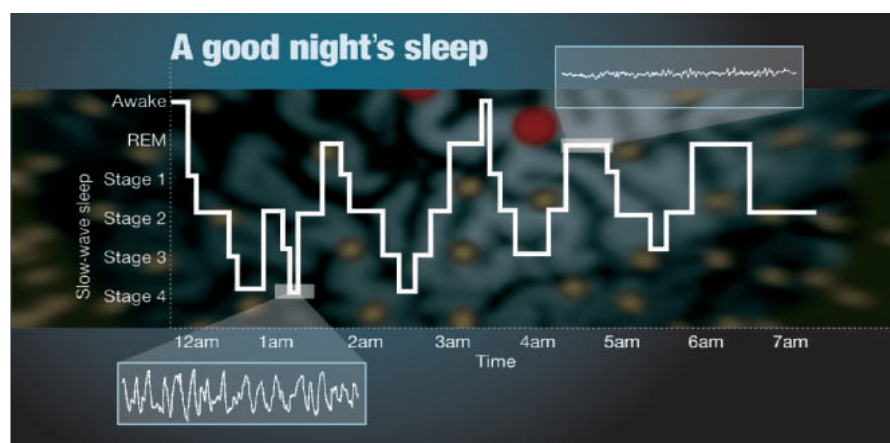
Inspired in part by Tononi's work, Stickgold and his colleagues now want to investigate what happens in the brain as it strengthens memories and integrates them with one another. As a first step, Matthew Walker, one of Stickgold's Harvard colleagues, aims to vary the tasks that his volunteers perform before and after sleep, trying to determine which brain region is needed for each task using functional magnetic resonance imaging scans.

Walker is also interested in working out whether different types of slow-wave sleep are involved in different types of learning. Slow-wave sleep is composed of four distinct stages, beginning with initial drowsiness and becoming progressively deeper. Walker suspects that different stages might be more important for different types of procedural learning, such as how to play a piano or how to drive a car. "These are subtly different types of learning," he says.

Walker and Stickgold have already shown that the sleep-induced boost in performance on a learned typing task correlates with the amount of 'stage 2' slow-wave sleep, particularly towards the end of the night⁶.

The Harvard researchers are also interested in investigating the role of particular EEG traces, known as 'spindles', which occur near the onset of sleep. These have been suggested to facilitate long-term changes in networks of brain cells, and so could be involved in learning⁷. Each spindle lasts for 1–2 seconds, and consists of waxing and waning waves at a frequency of 7–14 hertz.

After this initial activity, a full night's sleep involves several bouts of REM, interspersed with periods of slow-wave sleep,



with each cycle lasting about 90 minutes (see graphic, above). Intriguingly, it seems that — at least for learning to discriminate between different visual patterns shown on a computer screen — a nap of about 90 minutes containing all the phases of sleep produces about the same gain in performance as eight hours' sleep⁸.

Sleepy genes

As well as changes in electrical activity, memory consolidation during sleep may involve shifts in gene expression. Tononi's group has found that, in rats, expression levels of about 100 genes increase during sleep, independently of the time of day⁹. Some of these genes are already known to be involved in the release of neurotransmitter chemicals, through which brain cells communicate with one another.

These findings fit with the theory that sleep is involved in consolidating procedural memory. The levels of different neurotransmitters in the brain fluctuate during the different phases of sleep. And Stickgold has unpublished data which show that cocaine addicts and schizophrenics, who both have dysfunctional neurotransmitter systems, do not show normal sleep-

dependent procedural learning.

Establishing the precise relationship between sleep-induced procedural learning and the underlying changes in gene expression, neurotransmitter mechanisms and patterns of electrical activity will require extensive work in both human volunteers and experimental animals. No one is expecting sleep to give up its remaining mysteries without a struggle. But if researchers in the field ever get stuck in interpreting the results of a particularly perplexing experiment, they at least know to follow Mendeleev's example: just sleep on it.

Laura Nelson recently completed an internship with Nature.

1. Siegel, J. M. *Science* **294**, 1058–1063 (2001).
2. Huber, R., Ghilardi, M. F., Massimini, M. & Tononi, G. *Nature* **430**, 78–81 (2004).
3. Hairston, I. S. & Knight, R. T. *Nature* **430**, 27–28 (2004).
4. Wagner, U., Gais, S., Haider, H., Verleger, R. & Born, J. *Nature* **427**, 352–355 (2004).
5. Walker, M. P., Brakefield, T., Hobson, J. A. & Stickgold, R. *Nature* **425**, 616–620 (2003).
6. Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A. & Stickgold, R. *Neuron* **35**, 205–211 (2002).
7. Sejnowski, T. J. & Destexhe, A. *Brain Res.* **886**, 208–223 (2000).
8. Mednick, S., Nakayama, K. & Stickgold, R. *Nature Neurosci.* **6**, 697–698 (2003).
9. Cirelli, C., Gutierrez, C. M. & Tononi, G. *Neuron* **41**, 35–43 (2004).
10. Rattenborg, N. C., Lima, S. L. & Amlaner, C. J. *Behav. Brain Res.* **105**, 163–172 (1999).
11. Aserinsky, E. & Kleitman, N. *Science* **118**, 273–274 (1953).

You don't need a licence (or PhD) to use your brain

Do scientists look down on the general public with whom they seek better relations?

Sir—Enough. I cannot take it any more. I became so incensed after reading Adam G. Hart's letter "If you can lose a driving licence, why not a PhD?" (*Nature* **430**, 503; 2004) that I could not continue until I had replied. You see, I am part of the "general public" to which reference is so often made. Because I do not hold a PhD, I do not have a "licence to perform science". Therefore I should not be allowed to form hypotheses nor to gather data in support or refutation of them, although this seems like an everyday experience. Perhaps I should not be allowed a subscription to *Nature*.

Every issue I read of *Nature* (and of *Science*) contains at least one news article, book review, commentary or

letter that makes a distinction between scientists and non-scientists: presumably those with a PhD and those without. The implication is that possession of a PhD somehow makes a person different from, and better than, those of us who have chosen a different career path. It is as if the lack of a degree prohibits people from being knowledgeable or from thinking critically.

Contrary to Hart's assertion, and as a member of the general public, I do not perceive "holders of the degree" as being experts in their field — no more so than I would automatically assume that a licensed plumber or lawyer is necessarily competent and honest. Holding a law degree does not, by itself,

allow an individual to practise law.

There are constant calls for support from, better communications with, and understanding from, the general public. To an infrared astronomer, a molecular biologist is part of this general public, even though both have advanced degrees. The PhD versus non-PhD dichotomy is condescending and very poor "general public" relations.

Science is a philosophy, a problem-solving methodology, a way of looking at the world that has nothing to do with a specific institution's requirements for a degree.

David L. Anderson

13609 E. Cornell Avenue, Aurora,
Colorado 80014, USA

NIDA's commitment to tackling drug abuse

Sir—Your News Feature "A hard habit to break" (*Nature* **430**, 394–395; 2004) reports on the commitment of Nora Volkow, the director of the National Institute on Drug Abuse (NIDA), to pursue science despite pressures to follow politically motivated agendas.

However, the article less fully describes the mission, goals or scientific achievements of NIDA over the past 30 years. NIDA's previous director, Alan Leshner, should be credited for conveying a clear science-based message that drug addiction is a brain disease, deserving the same research and medical-insurance support as other diseases targeted by the National Institutes of Health (NIH).

Contrary to the assertion that NIDA shuns research into 'harm reduction' strategies, NIDA studies have examined the impacts of needle-exchange programmes and the effectiveness of office-based methadone maintenance — with documented success. NIDA has also funded extensive research on methadone to understand exactly what features of methadone maintenance make it an effective treatment for opiate addiction.

NIDA has supported basic and applied clinical research on the molecular, neurobiological and behavioural basis of drug abuse and addiction. It has also supported neurochemical, pharmacological and behavioural research for development of long-term treatments for each of the specific addictions. NIDA has funded research on the treatment, epidemiology and prevention of drug abuse, in accord

with its NIH research mandate.

Like other NIH institutes, NIDA does not set policy. Rather, it supports the best science, providing information, which, if appropriately used, can inform the development of enlightened policies. NIDA supports research that documents the most effective approaches for treatment of a disease. Unlike the diseases and diverse disorders that are the domain of other NIH institutes, drug abuse and drug addiction have only recently been accepted as diseases of the brain with molecular genetic, environmental and drug-induced bases. As such, these diseases, like all others, require appropriate intervention and chronic treatment for some, and prevention for others.

Mary Jeanne Kreek

The Rockefeller University, 1230 York Avenue,
New York, New York 10021, USA

Other signatories to this letter:

Reese Jones University of California, San Francisco, USA

Herbert D. Kleber Columbia University, USA

Thomas Kosten Yale University, USA

Charles O'Brien University of Pennsylvania, USA

Don't have a cow! Fight global warming with CFC

Sir—Human demand for meat and dairy products requires hordes of methane-spewing ruminants, the source of a major greenhouse gas. Therefore, the possible solution described in your News story "Vaccine targets gut reaction to calm livestock wind" (*Nature* **429**, 119; 2004) is exciting.

Yet, while we wait for a vaccine to be perfected, there is no excuse for wanton

methanogenesis. A chemical remedy has been known since 1967, when Thomas Bauchop of the University of California, Davis, found the answer (*J. Bacteriol* **94**, 171–175; 1967).

Bauchop collected the fore-stomach contents of cattle into flasks and studied methane production by the bacteria responsible. The work was fraught with troublesome foam, so an anti-foaming agent was sprayed into the flasks. The foaming stopped, but so did the methane production. It turned out that the aerosol propellant, a Freon (CCl₂F₂), inactivated methane-producing enzymes. This popular chlorofluorocarbon (CFC) was used as a coolant in refrigeration as well as an aerosol propellant, until its ozone-depleting activities were discovered during the 1970s.

Researchers at the University of Illinois determined that the Freon propellant attaches covalently to the cobalt ion of the vitamin B12 cofactor that is required by methane-yielding enzymes (J. M. Wood, F. S. Kennedy and R. S. Wolfe *Biochemistry* **7**, 1707–1713; 1968).

So, before the cinematically illustrated global-warming disaster of *The Day After Tomorrow* becomes reality, let us go forth and spray the cows with Freon.

Theodore A. Alston

Department of Anesthesia and Critical Care,
Massachusetts General Hospital,
Harvard Medical School, 55 Fruit Street,
Boston, Massachusetts 02114, USA

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

Growing biodiversity

Your local grocery store shows why variety matters.

Farmers' Bounty: Locating Crop Diversity in the Contemporary World

by Stephen B. Brush
Yale University Press: 2004. 320 pp.
\$37.50, £25

Stuart Pimm

Biodiversity's three-part definition — the variety of genes, species and ecosystems — is best appreciated on a beach, with tropical forest in front of you, coral reefs behind and mountains in the distance. On rainy midwinter days, in a city far away, there is always the local grocery. Upmarket ones, the habitat of yuppies, make the point; so do rural markets in poor countries. The former sell coffee from 20 countries, plus broccoli, kale, kohlrabi, cabbage and Brussels sprouts, which are all variants of *Brassica oleracea*. Rural markets in Central America will have variously coloured maize kernels of assorted sizes and types. In Andean valleys you may find a hundred kinds of potato. This is the biodiversity that *Farmers' Bounty* celebrates. It provides insights into questions of distribution, value and survival that apply to biodiversity as a whole.

Like species diversity, genetic diversity is concentrated in a few regions. These are known as Vavilov centres after a Soviet plant breeder of the 1920s and 1930s who was executed by Stalin for holding inconvenient scientific views. In *Farmers' Bounty*, Stephen Brush wanders bravely into theoretical ecology to find explanations for how this diversity persists. Agricultural experiences are not widely appreciated by theoreticians, yet they offer important insights. Ecologists debate the generality of the idea that heterogeneous environments promote diversity. Meanwhile, across the world, farmers discuss the advantages of this or that variety in different environments and years as seeds are sorted, and those retained are bought, sold, exchanged and mixed for next season's harvest.

In contrast, uniformity was the norm for potato-growers in Ireland in 1845 and US maize-growers in 1972, and the penalty was devastating. Maize-growers in modern Chiapas in Mexico must contend with risks from summer droughts, strong winds, various soil types and the uncertainties of labour for weeding or applying fertilizer. Growers understand the complexities of the relative advantages in taste, yield and storage of traditional varieties, improved ones and creolized varieties that are the result of decades-long mixing of the two, and decide the proportions of these that they will plant. Simply, crop diversity is the product of a



Rich pickings: markets in Kenya have a wide range of potatoes as farmers strive to improve crops.

global annual ecological experiment. It is unregulated and informal, changes constantly as varieties come and go, and is laden with tradition, language and myth. The participants' lives depend on their correctly interpreting the experiment's results.

Modern varieties, with their higher yields, better ability to use fertilizer and resistance to specific diseases or stresses such as drought, have been the major challenge to traditional diversity in the past half-century. 'New' potatoes are now grown in nearly every Andean valley. Governments encourage them, as does the grower's need for a larger surplus. Potato diversity has declined, as might be expected, but traditional potatoes persist, particularly at higher altitudes. They taste better, especially when boiled or baked, and make better gifts, facts not lost on the clientele at my local grocery.

Still, the erosion of variety can be substantial. The Chilean island of Chiloé had some 200 varieties of potatoes in the late 1920s, half that a decade later, and fewer than 40 by 1970. Brush accepts the general trend and the problems it raises, but his frustration over the lack of global documentation is well taken. The putative features of genetic erosion are that indigenous varieties have limited geographical distributions, they change little from year to year, and their variety decreases as modern varieties, and fertilizer and pesticide use, increase. Studies certainly suggest these outcomes, but often equivocally. Environment, economy and culture often conspire to conserve diversity for reasons that need broader synthesis and assessment.

The familiar solution to retaining diversity is the use of seed banks, such as the

International Rice Research Institute (IRRI) in the Philippines. Brush's view of such collections is that they are necessary but far from sufficient. The heroine botanist of *Jurassic Park* enthuses over a tree not seen for 65 million years, then gasps as she notices the sauropod eating it. The film imagines extinct species grown from their DNA, but is silent on how to teach dinosaurs which trees to eat. Similarly, seed banks store DNA but lose the complex relationships between the variety, its symbionts and pathogens, and the traditions of the farmers who knew how to plant, tend, harvest and store it. *In situ* conservation is essential; so is the selection that goes with it. Women in Rwanda select bean varieties, and in Nepal rice and chickpea varieties, that perform better than conventional crop breeders can manage.

If diversity has value, who owns it? In the book's best chapter, Brush discusses the increasing polarity in answers. Fears of 'biopiracy' are widespread in many tropical countries. Stories of how the rich have robbed the poor — and continue to do so — influence research-permit applications even for those of us who travel with only notebooks. The Convention on Biological Diversity has a strong theme that plant breeders' rights trump poor farmers' rights. Yet it is often the poorer country that benefits from exchanges — Vietnam gets almost all of its rice from lines developed from IRRI, compared with just a sixth for the United States.

Some countries favour outright protection, such as Ethiopia's ban on the export of coffee plants. Major coffee producers such as Colombia and Costa Rica have a narrow genetic base to their crops; diseases could

place them at Ethiopia's mercy. Mutual retaliatory actions between these countries would surely cause widespread harm.

Other countries have agreements for bio-prospecting, but these bring mixed benefits. It is not always clear, for example, who owns the intellectual property of the varieties. Brush's forceful conclusion is that increasing ownership will abuse the rights of people who have long been involved in the common pool of genetic resources, but who now find themselves excluded by modern patent laws. ■

Stuart Pimm is in the Nicholas School of the Environment and Earth Sciences, Duke University, Durham, North Carolina 27708, USA.

Myths and men

Moments of Truth: Four Creators of Modern Medicine

by Thomas Dormandy

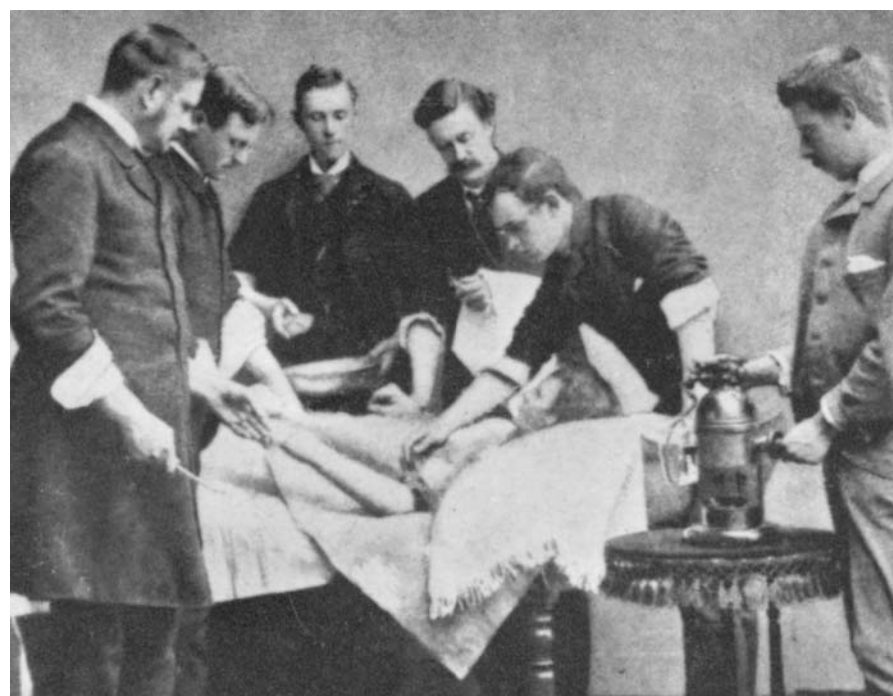
Wiley: 2004. 576 pp. £18.99, \$30

John Galloway

To misquote Jane Austen shamelessly: "It is a truth not sufficiently universally acknowledged that information is useful only when it provides evidence for or against some view or proposition." The "moments of truth" of Thomas Dormandy's title are the flashes of insight that propelled four nineteenth-century doctors into the medical pantheon. Théophile Laënnec invented the stethoscope; Ignác Semmelweis put an end to childbed fever; Joseph Lister is credited as the founder of antiseptic surgery; and Walter Reed proved that yellow fever was spread by mosquitoes, laying the foundation for its eradication.

The question here is whether Dormandy's book is just a chronicle of the lives of men whose greatness is already a given, or whether it is more analytical and critical, and makes any judgement on the nature and measure of their greatness. Were they justified in taking — or being given — personal credit for a particular breakthrough in medical science? Or were they each just one of several contributors who all deserved some recognition? I think it is fair to say that *Moments of Truth* has its feet firmly in the first camp. Dormandy says confidently of his four heroes: "They wrote their own papers and books and signed them without equivocation." They were not team players, then.

I enjoyed this book. The life-and-death business of medicine is well-nigh irresistible, as every fan of TV programmes such as *ER* or *Cardiac Arrest* knows. The medical background for Laënnec was tuberculosis, which



Medical mystery: Théophile Laënnec (below) certainly invented the stethoscope, but how much did Joseph Lister's carbolic acid spray contribute to early antiseptic surgery (above)?

killed thousands of people each year in France. Obstetric wards in Vienna, filled with poor pregnant women dying of puerperal fever, were the stamping ground for Semmelweis. Lister's province was the operating theatres of England and Scotland. As the great surgeon James Simpson said: "A patient laid on an operating table in one of our surgical

hospitals is exposed to more chances of death than the English soldier on the field of Waterloo." Reed was faced with recurring epidemics of yellow fever among the occupying US troops in Cuba, which killed 20 times as many soldiers as the war. Reed, a military pathologist, headed the commission charged with fighting the disease. For Reed, the larger historical context was crucial in putting him in the right place at the right time, with colleagues who already had a pretty shrewd idea that yellow fever was not directly contagious but spread by mosquitoes — all they lacked was proof.

The stories that Dormandy tells are good, but how do they stand up as history and biography? Novelists, playwrights and journalists recognize one thing above all others: that the essence of drama is that the part (of the hero) is greater than the whole. In real life, however, although individuals do great things, they rarely do them alone. Yet when scientific discovery is presented as drama, it must play down the contribution of others. It

is not that they are not mentioned, but rather that they play Rosencrantz and Guildenstern to the great man's Hamlet. We are not given the opportunity to know them as well as we know the central figure. How would the 'moment of truth' look through the lens of the biographies of these minor players, if they were to be written?

Dormandy picks up this point in the book's epilogue but asserts the opposite view, that great discoveries are made only by great individuals, and that no medical scientist illustrates this better than the four men in his book. "Long, long may they flourish," is how he signs off. This presents something of a problem. He decries John Macleod, who shared the Nobel prize with Fred Banting in 1923 for discovering insulin, apparently on the precept that rewarding the unworthy robs the deserving. "He had spent the time of the crucial experiment hunting in the Rockies, unaware that anything interesting was happening in his laboratory." In reality, Charles Best, who worked closely with Banting, was overlooked by the Nobel committee and spent the rest of his life campaigning for recognition. More significantly, arguably the most able member of the team, James Collip, seems to have been written out of the story altogether.

And what about Lister? Was he really the lone genius who ushered in the era of antiseptic surgery single-handedly? Did he then spend 40 years crying in the wilderness, before finally gaining the credit he so richly deserved? Dormandy marshals his facts to give a resounding yes. But some historians think otherwise. They say that the aseptic movement — with its clean wards, bedding

Science in culture

A measured approach

Alex Colville's exhaustive search for mathematical probity.

Martin Kemp

There is often a point in research of a mathematical nature when the solution to a problem is sensed before it is proved. An anticipatory instinct tells us when something feels right. For someone publishing a scientific paper, the proof is of course essential. For an artist who works towards pictorial solutions, the proof of the pudding is in the viewing, and the instinctual element of 'rightness' remains paramount.

Alex Colville, to my mind the best Canadian artist of his time, works with a slow-burning meditative persistence on the geometrical structures that underlie his pictures. The concept for the picture is triggered by his vision of a pregnant moment in the life of things. The picture shape that 'fits' the subject is then altered by repeated geometrical structuring using complex overlays from Colville's repertoire of inscribed circles, squares, triangles, logarithmic spirals and ratios. Human proportionality, which is based on the system of head lengths beloved of Renaissance architects and more recently Le Corbusier, has an important role in those paintings that involve his cast of wordless characters.

Series of carefully dated studies are made, looking like the storyboard for a film, but destined never to extend in time beyond one enigmatic still. Concentrated flurries of intense design are interspersed with contemplative intervals. Motifs are rearranged and manoeuvred in space, as Colville seeks, for both the mathematical structures and the subject, the right organization of space, the right surface conjunctions and the right intervals. These are neither wholly calculated nor freely instinctual, but a combination of both.

The process behind the picture shown here, *The Surveyor*, is fairly typical, involving almost 30 drawings over 14 months during 1999–2000. The surveyor's transit with its splayed tripod is measured for

precise transcription. A complex fretwork of lines drawn with straight-edge and compasses is intricately interwoven with the subjects, playing off those constituent parts that are most overtly geometrical with those whose order has been hidden behind the chaos of processes. Scales of head lengths are mapped on the compositional studies. Intricate sets of alternative diagonals structure the complex foreground region anchored by the legs of the tripod and surveyor. Tones and colours are positioned to see what works. All the while, neat calculations of scale keep everything in proportion.

The scene is in Nova Scotia, overlooking the tidal planes near Wolfville, where Colville lives. He is a local painter in the sense that Constable was local, creating art that has to draw nourishment from scenes known intimately in order to find a wider truth. The surveyor, in this case a woman, stands on a ribbon of isolated road, absorbed in the act of measuring. But measuring what? She seems to have turned the precise eye of her instrument on the elusive curves of the tidal waterway, which are always in the process of coming and going.

Colville's art is underpinned by his quest for

order from apparent disorder. He searches, like Piero della Francesca in the Renaissance or Georges Seurat in the late nineteenth century, for what we can find beneath and within the surface of appearances if we probe intensively enough. The visual reconstruction of order is hard won and fragile, like the order of society itself. For Colville, the quest, when conducted at the highest level, is a lonely one, drawing on deep resources, yet it is one we recognize and share.

The frame of Piero della Francesca's painting *Flagellation* (see *Nature* **390**, 128; 1997), in which three mute figures stand in the foreground, closely grouped and emotionally dislocated from the main subject, once bore the inscription, "*convenerunt in unum*", which roughly means "they conspire together". Colville's art is one of wordless conspiracies in which he enrolls us compellingly in his quiet narratives of the apparently ordinary. Below the surface are the strangeness and wonder, disorder and order, that have fired the quests of artists and scientists since the beginning of both art and science.

Martin Kemp is professor of the history of art at the University of Oxford, Oxford OX1 1PT, UK, and co-director of Wallace Kemp Artakt.



and operating clothes — long pre-dated Lister's spraying of open wounds with carbolic acid, which anyway turned out to be a cul-de-sac rather than the main highway of antiseptic surgery. John Waller has argued forcefully in his own book, *Fabulous Science* (Oxford University Press, 2002), that Lister was a tireless self-publicist, taking personal credit for a movement he did not start and to which he contributed relatively little.

So, are Dormandy's heroes giants stand-

ing on each other's shoulders, or dwarfs treading on each other's toes? Are their stories myth or history? In reality, there is no clear demarcation: it is all down to points of view and motives. After all, one of the functions of myths is to answer questions about origins. Where did medicine come from? Who invented antiseptic surgery? But that is also one of the functions of history. Consider the subtitle of this book, *Four Creators of Modern Medicine*. However painstakingly

collected and scrupulously verified, the real significance of historical 'facts' depends on the quality of the point of view for which they are marshalled. Of course, a point of view may be worth entertaining even if you do not actually accept it on the evidence, and that may depend on how it is put across. In this book, Dormandy puts his across rather well. ■

John Galloway is at the Eastman Dental Hospital, 256 Gray's Inn Road, London WC1X 8LD, UK.

Pathways and building blocks

Modularity in Development and Evolution

edited by Gerhard Schlosser & Günter P. Wagner

University of Chicago Press: 2004. 600 pp.
\$90, £63 (hbk); \$35, £24.50 (pbk)

Hans Meinhardt

Modules are relatively independent building blocks. This simple definition suggests that it should be relatively easy to recognize them in the development and evolution of biological organisms. However, there are many degrees of modularity, so any definition is likely to have some ambiguities. One way of identifying developmental modules is the repetitive use of a particular mechanism in different developmental situations. Another characteristic is a tight internal connection between the parts of the module, with fewer, but well defined, connections to the rest of the system in which the module works.

Investigations in development and evolution have been burdened by what Gerhard Schlosser and Günter Wagner refer to as “ugly facts”: a plethora of details that obstruct our view of the underlying principles. Their book is an attempt to overcome this problem by bringing a range of diverse fields together under a common theme. They have assembled more than 20 articles that discuss modularity at different levels: in gene regulation networks, in signalling pathways, in structures such as segments or appendages, and in symbiosis, for example. The authors are authorities in their respective fields and describe their subjects — such as helix-loop-helix proteins, signalling pathways, selector genes or individual versus colonial forms of life — in great detail.

The relevance of these descriptions to the central theme of modularity is not always spelled out, however, with only brief comments at the beginning or end of some of the chapters. To counteract this trend, Wagner and Schlosser have provided an opening introduction that seeks to integrate various themes, as well as separate introductions to each of the four main parts of the book. Promoting communication between different communities is certainly a difficult, but necessary, task.

An archetypal module in higher organisms is the cell. In the book, the transition from single cells to multicellular organisms is illustrated only by a rather specific example, the volvoclean green algae, which exhibits a smooth transition between life as single cells and multicellularity. However, the role of the cell as a module of development can hardly be overestimated. Many of



Poles apart: why do the legs and body of *Drosophila* have different segmentation systems?

the universal capabilities of extant higher organisms that we see in cells presumably evolved before the first multicellular organisms appeared. For example, all cells generate polarity. A pertinent but, at least in the discussion of signalling pathways, often neglected question concerns how modules first arose in evolution. One possibility is that a whole functional unit, including the genes that specified the components of the unit, was re-used in a modified form; the cell is a good example here. Alternatively, a module (such as a signalling system) might be co-opted for a function totally unrelated to its original purpose.

It is still difficult to find the origin of each process that we regard as a module, even with all the data that are now available. For example, the WNT signalling module, mentioned in several chapters of the book, is involved in the formation of the segment polarity network in the fruitfly *Drosophila*, in the generation of the primary anterior-posterior body axis in vertebrates, and in the generation of cell polarity. Do these different functions have a common root? The evolutionary origin of segmentation is still controversial, so we do not know whether there is a direct connection between the polarity of the axis of the whole body and the polarity of individual segments. Likewise, is there a connection between the determination of the axis of a single cell and the axis of a whole body? After all, the development of multicellular organisms starts with the polarization and division of a single cell. The few basic signalling pathways satisfy one criterion of being modules as they are used again and again. It would be interesting to know the

most ancestral organism in which these pathways have been clearly identified, and what their function was.

One chapter that I found particularly interesting deals with the segment polarity network in *Drosophila*, illustrating the importance of, and the difficulties encountered in understanding, the dynamics of a system; this cannot be derived from a list of the parts. The segment polarity network is not a module in the sense of being used repeatedly. The segmentation of the fly's legs is based on the *Notch* system, whereas the segmentation of the body depends on the *engrailed/wingless* pathway. This raises an interesting question for modularity: why was an existing network not redeployed, despite the fact that in both cases segments had to be formed?

The concept of modules spans many disparate fields of biology. This book provides a good introduction to the subject, despite the inevitable heterogeneity of a multi-author volume. Modularity goes to the heart of a central problem in biology: how has such an incredibly complex system as a higher organism evolved? A combinatorial assembly of elementary building blocks is certainly one way of generating complexity, especially if the building blocks were duplicated and modified during evolution. Many of these building blocks are carefully described in this book, which covers a much broader range of topics than most textbooks. Reading it should certainly sharpen your view of what are the interesting questions to ask. ■

Hans Meinhardt is at the Max-Planck-Institut für Entwicklungsbiologie, Spemannstrasse 37–39, 72076 Tübingen, Germany.

Going into reverse

Reversible computation: how feasible is a computer that is both logically and physically reversible?

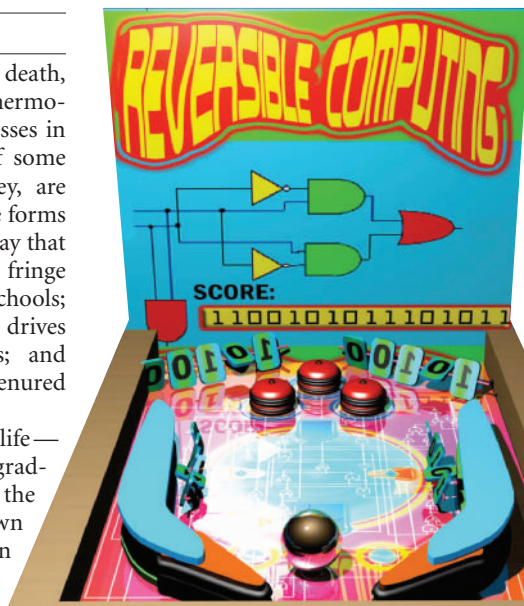
Seth Lloyd

Nothing in life is certain except death, taxes and the second law of thermodynamics. All three are processes in which useful or accessible forms of some quantity, such as energy or money, are transformed into useless, inaccessible forms of the same quantity. That is not to say that these three processes don't have fringe benefits: taxes pay for roads and schools; the second law of thermodynamics drives cars, computers and metabolisms; and death, at the very least, opens up tenured faculty positions.

Indeed, most of the good things in life — including life itself — arise from this gradual degradation of the useful into the useless, of order into disorder, known in physical terms as an increase in entropy. Still, wouldn't it be nice sometimes to slow down or even momentarily stop this inevitable dissipation of resources? After all, the second law of thermodynamics doesn't require that entropy always increase: it leaves open the possibility that entropy can remain constant. There are some processes that are effectively reversible: no energy is dissipated, and entropy remains constant, or almost so. Chemical reactions are reversible if run sufficiently slowly; they can dissipate arbitrarily small amounts of energy per step. Coherent quantum-mechanical processes such as tunnelling and superconductivity are reversible and can operate without dissipation. And so can computation, in principle.

More than 40 years ago, Rolf Landauer realized that there was a fundamental connection between computation and the second law of thermodynamics. Just like physical processes, logical processes can be either reversible or irreversible. A logic operation is reversible if its inputs can be inferred from its outputs. For example, the NOT operation that takes 0 to 1 and vice versa is reversible because if $Y = \text{NOT } X$, then knowing that Y is 0 allows one to deduce that X is 1. By contrast, the ERASE operation, which always gives the output 0 regardless of its input, is irreversible: the output gives no knowledge of the input.

Landauer pointed out a strong connection between logical irreversibility and physical irreversibility. First, logically reversible operations such as NOT can be implemented using physically reversible operations. For example, if a proton spinning clockwise around some axis is taken to register YES,



and one spinning anticlockwise registers NO, then application of a magnetic field causes the proton to flip from YES to NO and back again, all without dissipation or increase in entropy. The flipping spin implements a NOT operation in a physically reversible way. Second, a logically irreversible operation such as ERASE requires physical irreversibility. A bit in your computer's memory is registered by the charge in a tiny capacitor: when you erase the bit by restoring the capacitor to an uncharged state, then half the time the energy in the capacitor is dissipated as heat.

Conventional computations are built up of both reversible and irreversible logical operations, so at first the idea that logical irreversibility requires physical irreversibility (which is known as Landauer's principle) seemed to imply that computation requires dissipation. But then Charles Bennett and, independently, Ed Fredkin realized that most logically irreversible operations can be embedded in slightly more complicated reversible operations. As a consequence, computation can in principle be implemented using reversible physical processes such as slow chemical reactions or coherent quantum processes. The one logically irreversible process that cannot be embedded in a more complicated reversible process is erasure: when you erase the last copy of a bit from your computer, entropy must increase elsewhere.

In the past decade, logically and physically reversible computations have been implemented using logically reversible CMOS

(complementary metal oxide semiconductor) circuits. Perhaps even more impressive, prototype quantum computers, even though very small and simple, generate vanishingly small amounts of entropy at the microscopic level. Quantum computers are the closest thing we have to devices that carry out computation in a logically and physically reversible fashion. (Of course, the lasers, superconducting magnets and dilution refrigerators that are needed to run quantum computers generate plenty of heat. But the quantum logic operations that take place in the 'guts' of the quantum computer are themselves fundamentally reversible and dissipate almost no energy.)

What's the catch? In Bennett's original design and in reversible CMOS computers, the catch is that dissipation only goes to zero in the limit of zero speed. In the case of quantum computers, the catch, to paraphrase John Donne, is that no qubit is an island, entire unto itself. No matter how weak a qubit's interactions with the rest of the Universe may be, sooner or later one of them will flip that qubit and introduce an error into the computation.

Like added entropy, an error consists of unwanted information. Computers can be fitted out with error-correcting codes that seek out errors and correct them, by erasing the erroneous information. By Landauer's principle, erasure of a bit in the computer requires entropy increase elsewhere. For computers, error is the ultimate source of entropy.

Of course, slow computations full of errors can be performed with as little dissipation as you like. But the second law says that if you want to compute fast and accurately, you have to dissipate.

What about death? Must computation, like all physical processes, grind to a close? In fact, as long as the expansion of the Universe keeps on supplying free energy and environment into which to reject errors, the known laws of physics apparently allow a computer to compute for ever. Just what might an eternal computer compute? The answer will have to wait.

Seth Lloyd is in the Department of Mechanical Engineering, Massachusetts Institute of Technology 3-160, Cambridge, Massachusetts 02139, USA.

FURTHER READING

See articles by R. Landauer, C. H. Bennett and S. Lloyd in *Leff, H. S. & Rex, A. F. (eds) Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing*, 2nd edition (Institute of Physics, Bristol, 2003).

Cancer and the homeless cell

Lance A. Liotta and Elise Kohn

A protein has been identified that enables cells to survive when dislodged from their substrate, and to migrate to new sites in the body. Such a mechanism might give cancer cells a significant advantage.

Usually, the cells of an organ stay close to home. In fact, their lives depend on it. Within their own neighbourhood, they communicate to mutual benefit with the cells around them, while signals from the matrix beneath tell them that they are on home ground¹. When they lose contact with the matrix they die, in a process named 'anoikis' (from the Greek for 'homelessness')². By contrast, cancer cells do not become fatally homesick if separated from home turf — quite the opposite. The ability to leave home and survive, even thrive, in a foreign tissue is essential for them to invade and metastasize. This strategy promotes the expansion and dissemination of the cancer-cell population, but sadly is often fatal to the patient. Now Douma and colleagues³ describe how, by screening for cells that can resist anoikis, they have identified a new protein player in metastasis (page 1034 of this issue).

Douma *et al.*³ studied rat intestinal epithelial cells — those that line the intestine. The malignant transformation of epithelial cells is described classically as the development of anchorage-independent growth⁴. To identify genes that enable cells to grow independently of anchorage to a matrix — that is, to survive the threat of anoikis — Douma *et al.* carried out an unbiased genetic screen. They delivered DNA molecules containing different genes into the rat cells in culture, and then transferred the cells from adhesive culture dishes into a non-adhesive environment.

Control cells died when separated from their matrix in this way. But one group of treated cells survived. These, the authors discovered, had received DNA molecules encoding the TrkB protein. TrkB is best known for its role in the nervous system, where, together with its main ligand, brain-derived neurotrophic factor (BDNF), it promotes the proliferation, differentiation and survival of normal neural components such as retinal cells and glial cells⁵. The epithelial cells described by Douma *et al.* co-opted this mechanism to overcome the threat of death through homelessness.

But the authors also found that the detached TrkB-expressing cells survived and continued to grow as spheroid aggregates. So could they have survived simply because they replaced adhesion to their substratum with adhesion to neighbouring cells? Appar-

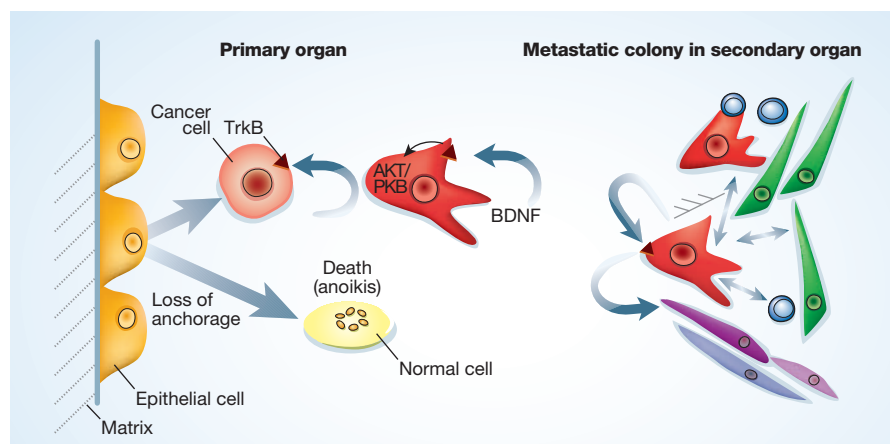


Figure 1 Survive and disseminate. Epithelial cells are normally attached both to each other and to their matrix. They undergo programmed cell death when detached from their home ground — a process called anoikis. In contrast, cancer cells can survive by using autocrine or paracrine mechanisms to suppress programmed cell death, stimulate tissue invasion, and promote the growth of new blood vessels to supply oxygen. As Douma *et al.*³ describe, the autocrine (autostimulatory) mechanisms include the production of TrkB protein, which is stimulated by brain-derived neurotrophic factor (BDNF) and in turn activates the AKT/PKB protein. Paracrine mechanisms involve interactions with external components (such as immune cells, other matrices, the cells of blood vessels, and so on). For simplicity, paracrine mechanisms are shown as operating only in metastatic sites, but in reality they can also function in primary tumours.

ently not. Further experiments by Douma *et al.* showed that, instead, TrkB production seemed to allow cells to resist anoikis by rewiring their internal circuitry. Specifically, TrkB triggered the activation of phosphatidylinositol-3-OH kinases (PI(3)Ks). These enzymes are known to stimulate another enzyme, the protein kinase AKT/PKB (ref. 6), and this too was active in the TrkB-producing cells. The result was a blockade of caspases, the DNA-cleaving enzymes that mediate anoikis and related forms of cell death.

PI(3)K enzymes also drive numerous other cellular functions associated with metastasis^{6–8}. For instance, the PI(3)K pathway enables the cellular 'skeleton' to deform and 'feet' to protrude, thereby allowing cells to move — a key aspect of tissue invasion. The pathway is also implicated in stimulating the production of new blood vessels within tumours when oxygen levels are low. Not surprisingly, deregulation of the PI(3)K pathway has been reported in all the major types of epithelial cancer, and can be caused by the loss of upstream suppressor proteins, the constitutive activation of PI(3)Ks or the activation of downstream elements — and, it now seems, by the production of an

upstream activator, TrkB. The involvement of this pathway may explain how TrkB by itself causes rat intestinal epithelial cells to be resistant to anoikis, and why, as Douma *et al.* show, this is sufficient to allow these cells to metastasize when injected into mice. *In vivo*, the inherent genetic instability of cancer cells might lead them to upregulate TrkB expression and activate downstream events.

There are several examples of molecules that are necessary, or even sufficient, to induce the full spectrum of characteristics necessary for metastasis in experimental models. A recent example is Twist, a gene-transcription factor that is better known for its regulation of changes in body shape during development⁹. We can speculate that the molecular programme required for metastasis can be triggered from a variety of starting points. But, given Douma *et al.*'s findings, we might also expect any process of metastasis induction to include a component that renders cells able to survive independently of attachment to the matrix or to other cells. In keeping with this expectation, forced expression of Twist stimulates cell invasion and causes the loss of cell–cell adhesion that is mediated through the E-cadherin protein⁹. Perhaps Twist can also co-opt the PI(3)K

pathway and AKT/PKB, which are known to be involved in E-cadherin-mediated cell–cell recognition¹⁰.

Given that TrkB enables cells to survive when homeless, and thus to metastasize, might blocking this protein be an effective anticancer treatment? Perhaps, but we should be realistic about this expectation. Tumour cells can survive by means of an autostimulatory (autocrine) signalling loop, such as that mediated by TrkB and BDNF, or through a paracrine cross-communication with their environment (Fig. 1). Such cross-talk might use the same molecules: for example, BDNF is produced by the cells that line blood vessels, and production is enhanced by low oxygen conditions¹¹; these events could promote the survival of both tumour cells and newly formed blood vessels, in the primary tumour or at a metastatic site. But there are other molecules involved in paracrine signalling, and we should be prepared for the possibility that strategies that target TrkB could be thwarted through these channels.

Homelessness is only one of a multitude of environmental stresses that can befall a cell. One way to overcome such problems is to seek a new and more salubrious location. Thus cells frequently need to migrate in order to survive. For instance, yeast that are poorly nourished extend migratory strands (hyphae) to seek better conditions. Similarly, slime-mould colonies that are deprived of nutrients send out scouts. And cancer cells, faced with a shortage of oxygen, hormones or nutrients, or with overcrowding, might survive by metastasizing — invading local tissue, entering the lymph and blood streams, exiting the circulation at a distant site, and, finally, establishing a secondary colony.

Avoiding death in a harsh or foreign environment seems to be a basic requirement for all types of nomadic behaviour. A mechanism such as that described by Douma *et al.*³, which promotes both survival and migration, would give tumour cells (and new blood vessels) a significant advantage. It might even be essential for tumour cells to pass the fitness test for metastasis.

Lance A. Liotta and Elise Kohn are at the National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-1500, USA.
e-mail: lance@helix.nih.gov

1. Liotta, L. A. & Kohn, E. *Nature* **411**, 375–379 (2001).
2. Frisch, S. M. & Ruoslahti, E. *Curr. Opin. Cell Biol.* **9**, 701–706 (1997).
3. Douma, S. *et al.* *Nature* **430**, 1034–1040 (2004).
4. Lawlor, E. R., Scheel, C., Irving, J. & Sorensen, P. H. *Oncogene* **21**, 307–318 (2002).
5. Huang, E. J. & Reichardt, L. F. *Annu. Rev. Biochem.* **72**, 609–642 (2003).
6. Wendel, H. G. *et al.* *Nature* **428**, 332–337 (2004).
7. Brader, S. & Eccles, S. *Tumour* **90**, 2–8 (2004).
8. Davies, M. A. *et al.* *Cancer Res.* **58**, 5285–5290 (1998).
9. Yang, J. *et al.* *Cell* **117**, 927–939 (2004).
10. Pece, S. *et al.* *J. Biol. Chem.* **274**, 19347–19351 (1999).
11. Kim, H., Li, Q., Hempstead, B. L. & Madri, J. A. *J. Biol. Chem.* **279**, 33538–33546 (2004).

Materials science

Silicon carbide in contention

Roland Madar

Silicon carbide is a highly desirable material for high-power electronic devices — more desirable even than silicon. And now the problem of producing large, pure wafers of the carbide could be solved.

After a period of intense scientific and industrial development, the semiconductor silicon carbide (SiC) is at last proving capable of outperforming silicon in electronic devices for high-power, high-frequency and high-temperature applications. Its potential as a replacement for silicon has been known since the 1950s, but its late emergence is principally due to the difficulty of growing large SiC crystals of sufficient quality. That difficulty is set to evaporate, thanks to the crystal-growing process introduced by Nakamura *et al.*¹ on page 1009 of this issue.

A major problem encountered in growing crystals of SiC is that the material doesn't have a liquid form. This means that the traditional methods of crystal growth developed for silicon and other semiconductors, based on controlled solidification from the liquid phase, cannot be used for SiC. However, the invention in 1978 of the 'modified Lely method'² (or physical vapour transport, PVT) opened the way to the production of large-area SiC wafers. This is a seeded sublimation technique: supersaturated SiC vapour condenses onto a single crystal seed inside a graphite crucible. Impressive progress has been made, but the quality of the SiC grown in this way is still too poor and remains an obstacle to be overcome if SiC technology is to make a commercial breakthrough. Improving the structural properties and at the same time increasing the size of SiC wafers that can be grown are key areas of research in this field.

As well as the classic defects, such as different kinds of dislocation, that are common

to most materials, SiC wafers produced by PVT tend to have some peculiar defects of their own. Known as 'micropipes', these defects are a consequence of the crystal structure of SiC and profoundly affect the reliability of electronic devices based on this material. The basic unit of the crystal structure is a silicon–carbon bilayer; within each bilayer, carbon atoms are arranged in a two-dimensional hexagonal pattern, with silicon atoms directly on top of them. Three-dimensional lattices are then built by stacking the bilayers along the normal direction, called the *c*-axis, of the hexagonal structure, and a surface polarity effect results, with one face of the crystal formed of silicon atoms and the other of carbon atoms (Fig. 1). In fact, several crystal structures, or polytypes, are possible — all described in terms of stacked hexagonal layers, but with different rotational offsets between the layers. Micropipes are

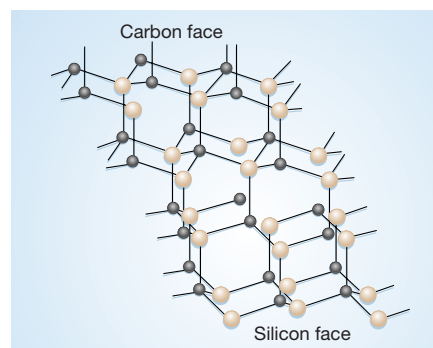


Figure 1 The structure of silicon carbide. The three-dimensional lattice is composed of hexagonally patterned bilayers of carbon (black) and silicon (white) in this 4H polytype. The termination of the crystal structure at a surface results in a face of carbon atoms or of silicon atoms.

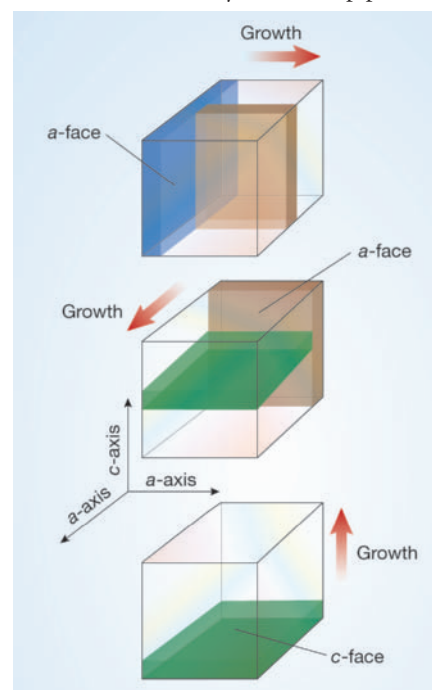


Figure 2 The repeated *a*-face growth process. In the first stage, silicon carbide (SiC) is grown on the *a*-face of a seed crystal. A segment of the newly grown crystal then becomes the *a*-face seed of the next growth stage. A slice of that growth seeds the final stage, but now the new crystal is grown on the *c*-face. The inventors of the process, Nakamura *et al.*¹, say that repeating the step of *a*-face growth before the more usual *c*-face stage eliminates faults in the resulting crystal.

hexagonal tube-like cavities, with diameters in the sub-micrometre to few-micrometre range, that develop parallel to the *c*-axis of the hexagonal structure (which is the common direction of crystal growth). They are the most harmful defects in SiC.

Intensive work on various aspects of the PVT growth technique — reactor design, growth conditions, seed orientation and surface preparation — has led to a considerable reduction in the number of these defects in SiC ingots and wafers. The best commercially available 3-inch-diameter wafers (of the 4H polytype) have micropipe densities of $10\text{--}100\text{ cm}^{-2}$ and dislocation densities of $10^3\text{--}10^4\text{ cm}^{-2}$. This quality of crystal is good enough for use in some commercial SiC devices, such as Schottky diodes³, but not for others. For instance, in high-power bipolar devices, there is a degradation in the material's electrical properties that seems to be related to the development of extended stacking faults, originating from in-plane dislocations in the SiC (ref. 4).

So reducing the density of dislocations in an SiC wafer, and in the epitaxial layer on top of it that forms the active part of the device, is an absolute necessity for the development of high-power SiC technology. Nakamura *et al.*¹ have succeeded: their new process produces SiC crystals with a dramatically lower dislocation density — although, perversely, the first step of the process actually creates a high density of stacking faults.

Most SiC crystals are grown on the *c*-faces

(perpendicular to the *c*-axis) of other crystals. In fact, it has already been shown that growing SiC crystals on surfaces with a different crystal orientation — the *a*-faces — fully suppresses the formation of micropipes. However, crystals made in this way are affected by basal-plane stacking faults. Nevertheless, Nakamura *et al.* started with a SiC single crystal that had been grown on an *a*-face, inheriting a high density of dislocations from its seed crystal. Then, taking a section of the crystal along the *a*-axis of that face, they allowed the crystal to develop on the other *a*-face (Fig. 2), afterwards continuing with classic growth on the *c*-face. Nakamura *et al.* call this the repeated *a*-face (RAF) growth process. It is the repetition of the *a*-face step that ensures that stacking faults are eliminated and dislocations suppressed. The ingots of SiC produced are, say the authors, “virtually dislocation-free”.

These results are spectacular: the RAF process is a major innovation in materials science. Silicon carbide has become, at last, a contender for silicon's crown.

Roland Madar, of CNRS, is in the Department of Physics, Institut National Polytechnique de Grenoble, Grenoble 38402, France.
e-mail: roland.madar@inpg.fr

1. Nakamura, D. *et al.* *Nature* **430**, 1009–1012 (2004).
2. Tairov, Yu. M. & Tsvetkov, V. F. *J. Cryst. Growth* **43**, 209–212 (1978).
3. Peters, D., Friedrichs, P., Schörner, R. & Stephani, D. *Mater. Sci. Forum* **389–393**, 1125–1128 (2002).
4. Lendenmann, H., Dahlquist, E., Bergman, J. P., Bleichner, H. & Hallin, C. *Mater. Sci. Forum* **389–393**, 1259–1264 (2002).

Social evolution

Kinship is relative

David C. Queller

Kinship fosters the evolution of cooperation. However, a once-heretical theory and an unconventional social organism show that the cooperation-enhancing effect of kinship is sometimes negated.

As evolutionary biologist W. D. Hamilton showed more than 40 years ago, selfish genes can lead to cooperation and altruism¹. A gene can spread by helping other individuals that carry copies of the gene. This extension of individual selection, called kin selection, works best with the recognition of close relatives, but Hamilton also thought that if the dispersal of individuals is limited, this might build up enough local genetic similarity to favour a less targeted kind of altruism towards neighbours in general. Because local dispersal is very common, this mechanism might greatly expand the range of cooperation in nature. However, later models^{2–4} suggest that this is not necessarily so: when limited dispersal makes neighbours close relatives, it also makes them close competitors, and this can negate the effect of kinship. On page 1024

of this issue⁵, Griffin and colleagues provide the first experimental test of this effect, confirming that both kinship and the scale of competition matter.

The study is part of a growing trend towards using microorganisms to study social evolution⁶. Sociobiologists have been slow to recognize that microorganisms have social interactions worthy of study. However, any reader of John Bonner's work on cellular slime moulds⁷ will realize that these single-celled amoebae interact in interesting ways; some give up their lives to form a stalk that furthers the dispersal of others. Analogous behaviour occurs in *Myxobacteria*⁸. And interesting behaviour is not limited to the exotica of the microbial world; the old laboratory workhorse *Escherichia coli* has strains that aid their own type by wielding poison–antidote systems to destroy other

strains that lack the same system⁹. Perhaps the most common form of microbial cooperation is the secretion of substances outside the cell to do various kinds of work, such as digestion. The work provides a public good, in which the benefits are available not just to the secretor, but to anyone in the neighbourhood. This puts individuals into a social bind. Why work to help others when you can be lazy and freeloader off your neighbours? The obvious answer is kin selection. If your effort helps others with the same genes, it is not wasted.

Once it is recognized that microbes have interesting social behaviour, two clear advantages can be exploited. First, many microbes are well understood biochemically and genetically. Second, it is easy to manipulate entire populations of microbes over space and to select them over time. Griffin *et al.*⁵ made full use of these advantages in their study of a freeloading mutant of the bacterium *Pseudomonas aeruginosa*. This ‘cheater’ mutant is deficient in production of a public benefit called siderophores. Siderophores are synthesized and excreted in response to iron deficiency. They then bind iron, making it available for uptake — either by the secretor or by its neighbours. The mutant saves on the cost of producing siderophores but can still obtain iron if it has secretor neighbours.

Griffin *et al.* selected populations, starting with equal numbers of cheaters and secretors, over six cycles of group interaction, each lasting about seven generations (Fig. 1, overleaf). In natural populations with limited dispersal, neighbours tend to be both relatives and competitors, but Griffin *et al.* were able to vary the relatedness and scale of competition independently. High or low relatedness was imposed by allowing bacteria to grow and interact in groups derived from a single bacterium (a cheater or a secretor) or from two bacteria (initially in a ratio of one cheater to one secretor), respectively. The scale of competition was manipulated during the transfers to form each next cycle of groups. Global competition was allowed by mixing the groups and then choosing random individuals to start the next cycle. The more productive groups thus contributed more to the next cycle. Local competition was imposed by taking an equal number of individuals from each group, without mixing between the cycles. Crossing the relatedness and competition treatments yields a simple two-by-two design that permits the study of both factors and their interaction.

As expected, at the end of the experiment the high-relatedness treatments yielded higher frequencies of the cooperative secretors than the low-relatedness treatments, demonstrating the importance of kinship in the evolution of cooperation. But the greater novelty of the study comes from another theory—

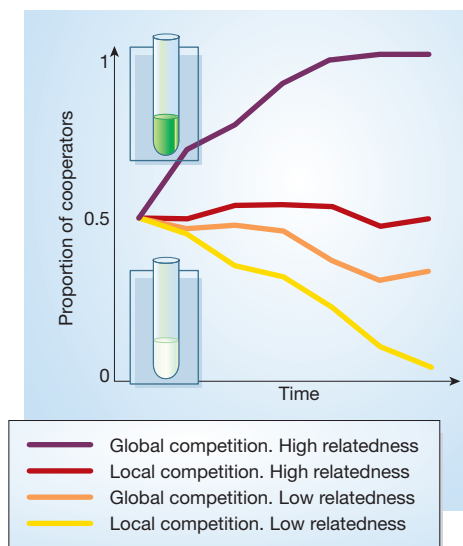


Figure 1 Cheaters and secretors. Cooperator *Pseudomonas aeruginosa* secrete a green siderophore, pyoverdine, which makes iron available for uptake by all cells. Cheaters save on the expense of producing pyoverdine but can exploit the siderophores of nearby secretors. Griffin *et al.*⁵ tested how the proportion of each varied over time in response to relatedness (high or low) and scale of competition (global or local). As these results show, relatedness is crucial but it is modulated by the scale of competition. (Simplified from ref. 5.)

confirming result: the local-competition treatments yielded more cheaters than the global-competition ones. Kinship is crucial, but it is modulated by the scale of competition. Local competition exactly cancelled out the effect of high relatedness, such that neither type was favoured (Fig. 1).

There is an interesting history behind the theories tested here. The first study to challenge Hamilton's ideas about limited dispersal was a computer simulation based on group-selection theory². This theory divides selection into within-group and between-group components, with the latter favouring cooperation. Group selection has a shady past. At one time, some biologists saw group-selected cooperation everywhere in nature, even though the process was unsupported by rigorous models, and consequently the whole idea was brought into disrepute. But there has been growing agreement that group selection, properly applied, is a roughly equivalent way of understanding social selection — not so much an alternative theory as an alternative means of slicing up the way in which selection occurs. Thus, kin-selection theorists took the computer-simulation results² seriously and quickly found two different ways of squaring them with their theory. First, one can add in the indirect effects for the cases where helping a neighbour takes away fitness from another, perhaps equally related, neighbour³. Alternatively, relatedness can be rescaled so that it measures genetic similarity

relative to the locally competing individuals, rather than to the global population⁴.

It is the relativistic effect of kinship that Griffin *et al.* set out to show, and they have succeeded. Curiously, however, their experiment is perhaps more easily understood from a group-selection standpoint. The conditions of low and high relatedness correspond exactly to the presence and absence of within-group selection. The conditions of global and local competition correspond exactly to the presence and absence of between-group selection. The two-by-two crossing of these treatments therefore leads to the most basic group-selection experiment possible. The results confirm that cooperation is favoured by between-group selection and disfavoured by within-group selection.

Does all this mean that we should discard kin selection in favour of the simpler group-selection approach? Hardly. Kin selection has yielded far more insights into the complex behaviours of animals such as social insects. For example, elegant theories of sex-ratio conflict in social insects emerge naturally from kin-selection models, whereas the corresponding group-selection models

are so complex that they have not been developed. But this history does suggest that group selection has matured enough for it to become a partner with kin selection in contributing insights, though the story is not without some irony. Once, group selectionists saw cooperation everywhere but were brought down to earth by individual selectionists. Now group selection is being used, not to show the ubiquity of cooperation but to rein in theories on an important form of cooperation envisaged by individual selectionists.

David C. Queller is in the Department of Ecology and Evolutionary Biology, Rice University, Houston, Texas 77005, USA.

e-mail: queller@rice.edu

1. Hamilton, W. D. *J. Theor. Biol.* **7**, 1–52 (1964).
2. Wilson, D. S., Pollock, G. B. & Dugatkin, L. A. *Evol. Ecol.* **6**, 331–341 (1992).
3. Taylor, P. D. *Evol. Ecol.* **6**, 352–356 (1992).
4. Queller, D. *Evol. Ecol.* **8**, 70–73 (1994).
5. Griffin, A. S., West, S. A. & Buckling, A. *Nature* **430**, 1024–1027 (2004).
6. Crespi, B. J. *Trends Ecol. Evol.* **16**, 178–183 (2001).
7. Bonner, J. T. *The Cellular Slime Molds* (Princeton Univ. Press, 1967).
8. Velicer, G. J., Kroos, L. & Lenski, R. E. *Proc. Natl. Acad. Sci. USA* **95**, 12376–12380 (1998).
9. Riley, M. S. & Gordon, D. M. *Trends Microbiol.* **7**, 129–133 (1999).

Chemical biology

Green fluorescent RNA

Michael Famulok

The future for intracellular imaging looks bright with the development of fluorescent probes made entirely of RNA. The cunning design exploits structural attributes of RNA to detect a variety of small molecules.

Green fluorescent protein (GFP) has proven tremendously useful as a method of creating intense visible fluorescence by entirely molecular biological means¹. Generally, the gene encoding GFP is fused with a gene of interest so that the resultant protein is tagged with the fluorescent module. It can then be followed in real time by optical imaging, to decipher spatio-temporal information on gene expression or protein localization within living cells or tissues². Such information is crucial for understanding complex cellular processes that depend on when, where and how much of a protein is present. The challenge now is to devise similar methods to reveal other molecules involved in cellular functions, such as small-molecule messengers, drugs, metabolites and short functional RNAs. Stojanovic and Kolpashchikov³, writing in the *Journal of the American Chemical Society*, report a considerable advance in this area with the development of RNA-based probes that can detect small organic molecules by fluorescence.

The new concept relies on aptamers — short RNA sequences that can bind specifi-

cally to particular ligands⁴. These molecules are fairly recent discoveries, but already they have been adapted to make probes for various molecules⁵. However, current aptamer probes are generated and labelled with fluorescence or radioactivity outside the cell, and a delivery mechanism is needed to get them inside — a procedure that is potentially disruptive to the very processes under observation. Stojanovic and Kolpashchikov have sidestepped this problem by creating an aptamer probe that can be genetically encoded and produced inside the cell, and that is able to generate marked fluorescence by itself.

The basis of their probe is an aptamer that can bind to the dye malachite green⁶ (MG). The dye itself is not fluorescent, but its structure is such that when it binds to the aptamer it is forced into a rigid conformation that is highly fluorescent⁷. This, however, is only half the story. Stojanovic and Kolpashchikov's probe has a second module — another aptamer that binds to whatever small molecule is to be detected.

To form the final probe, the authors take advantage of an intrinsic property of

aptamers: adaptive binding of ligands. Ever since the first nuclear magnetic resonance structures of aptamer–ligand complexes, it has been clear that the binding of a ligand by an aptamer occurs almost exclusively by adaptive recognition⁴. For example, aptamers often contain unpaired loop or bulge regions, which are disordered in the free nucleic acid and only acquire a defined conformation by adaptive folding around the ligand.

In the new probe, the MG aptamer and the detector aptamer are linked by a communication module. This module substitutes for one of the helices in the MG-aptamer domain, the formation of which is absolutely required for binding to the dye. When a ligand binds to the detector unit, adaptive binding forces the communication module into a helical structure, providing the right conformation for the MG aptamer to bind to the dye. The function of the short communication element is thus to render the disparate RNA domains interdependent, such that the probe can bind only to MG — and hence generate fluorescence — once a ligand has bound to the detection module (Fig. 1). It also means that the final sensor probe consists entirely of RNA and can therefore be generated by genetic encoding, requiring only the exogenous addition of the dye molecule to the growth medium to become an intracellular sensor.

In this respect, the aptamer probes are actually more analogous to another important genetic reporter system — firefly luciferase — than to the autofluorescent GFP. Luciferase requires the addition of a small-molecule substrate, luciferin, to function as a light-emitting reporter for the presence of a protein. The beauty of the new aptamer system is that it promises to have many of the advantages of GFP and luciferase, particularly in that the cells under investigation should be perturbed only minimally. The probes are less bulky than either GFP or luciferase, and could be designed for expression in cells without requiring covalent attachment to any protein, unlike GFP or luciferase. In addition, their modular design could potentially be useful for the intracellular detection of a large variety of molecules.

So far, however, the new aptamer probes are merely a proof of principle. To permit widespread cellular application, they still require optimization and improvement with regard to more straightforward design, choice of the chromophore, and sensitivity. Stojanovic and Kolpashchikov's first set of probes use known aptamers that bind to adenosine triphosphate⁸, theophylline⁹ or flavin mononucleotide¹⁰ to detect the molecules *in vitro*. Reporters that can trace compounds such as these are highly desirable because small-molecule messengers, such as Ca²⁺ ions, cyclic nucleotide monophosphates and lipids, often act locally and

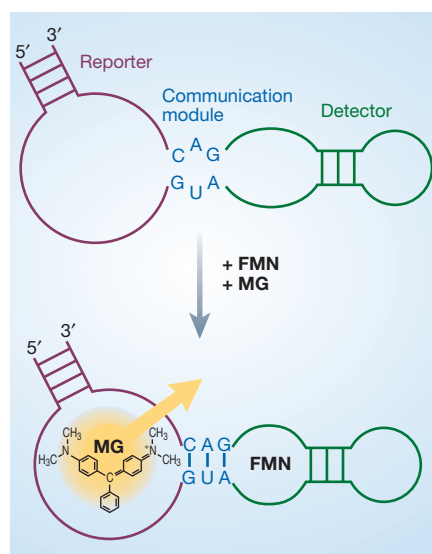


Figure 1 Fluorescent RNA-based probes. The probes designed by Stojanovic and Kolpashchikov³ consist of a reporter module that binds to the dye malachite green (MG), a communication module and a detector module, which in this example binds to flavin mononucleotide (FMN). The binding of FMN to the detector unit induces helix formation in the communication module. This helix completes the structure required to bind to the dye, which fluoresces only once it is bound to the probe.

transiently to control cellular functions, and systems to follow them are scarce².

It is possible that the MG-aptamer module could also be developed for the visualization of functional nucleic acids. Indeed, the fact that it is entirely based on nucleic acid would seem to predestine it for that purpose. Cells can silence certain genes at the stage

of translation by a mechanism that involves short single-stranded RNAs known as miRNAs. This process allows protein expression levels to be altered very quickly¹¹. The detection of miRNAs in live cells and tissues will help to address questions about their genetics, biogenesis, trafficking and function. Progress is already being made towards this end: recent work from my group used a design based on an RNA enzyme to develop fluorescent probes for specific miRNAs¹².

Together these studies highlight the importance of developing versatile methods for revealing the myriad cellular molecules as they move and change over time. Future endeavours in this area promise to revolutionize our insight into the fascinating interplay between the compounds that keep cells up and running.

Michael Famulok is at the Kekulé Institut für Organische Chemie und Biochemie, University of Bonn, Gerhard-Domagk-Strasse 1, 53121 Bonn, Germany.
e-mail: m.famulok@uni-bonn.de

1. Tsien, R. Y. *Annu. Rev. Biochem.* **67**, 509–544 (1998).
2. Zhang, J., Campbell, R. E., Ting, A. Y. & Tsien, R. Y. *Nature Rev. Mol. Cell Biol.* **3**, 906–918 (2002).
3. Stojanovic, M. N. & Kolpashchikov, D. M. *J. Am. Chem. Soc.* **126**, 9266–9270 (2004).
4. Hermann, T. & Patel, D. J. *Science* **287**, 820–825 (2000).
5. Silverman, S. K. *RNA* **9**, 377–383 (2003).
6. Grate, D. & Wilson, C. *Proc. Natl Acad. Sci. USA* **96**, 6131–6136 (1999).
7. Babendure, J. R., Adams, S. R. & Tsien, R. Y. *J. Am. Chem. Soc.* **125**, 14716–14717 (2003).
8. Huizenga, D. E. & Szostak, J. W. *Biochemistry* **34**, 656–665 (1995).
9. Jenison, R. D., Gill, S. C., Pardi, A. & Polisky, B. *Science* **263**, 1425–1429 (1994).
10. Burgstaller, P. & Famulok, M. *Angew. Chem. Int. Edn* **33**, 1084–1087 (1994).
11. Bartel, D. P. & Chen, C. Z. *Nature Rev. Genet.* **5**, 396–400 (2004).
12. Hartig, J. S., Grüne, I., Najafi-Shoushtari, S. H. & Famulok, M. *J. Am. Chem. Soc.* **126**, 722–723 (2004).

Transmissible spongiform encephalopathies

Prion proof in progress

Herman K. Edskes and Reed B. Wickner

Whether a protein can transmit disease in mammals has been an open question for some time. The latest test of this idea provides some strong evidence in favour, but is unlikely to end the debate.

Studies of the infectious agent responsible for the transmissible spongiform encephalopathies have spanned more than six decades. But the slow course of these diseases, and experimental difficulties in their study, has left the nature of this agent in doubt. Now, a report in *Science* by Legname *et al.*¹ represents a major step towards proving that a protein is the only essential element of the infectious agent.

Transmissible spongiform encephalopathies (TSEs) are uniformly fatal neurodegenerative diseases, and include scrapie in sheep, Creutzfeldt–Jakob disease (CJD) in humans, bovine spongiform encephalopathy

(BSE) and the latter condition transmitted to humans (variant CJD). The discovery that the infectious agent that causes scrapie is far more radiation-resistant than known viruses or bacteria — which rely on the very radiation-sensitive DNA or RNA — led to the suggestion² that this agent is a protein (without a required nucleic acid). Proteins are far less sensitive to radiation damage than are DNA and RNA. With the genetic³ and biochemical^{4–7} identification of PrP, a protein essential for the propagation and infectivity of TSEs⁸, the ‘protein-only’ hypothesis^{2,9} took concrete form (‘prion’ means ‘infectious protein’).

The protein PrP is a cell-surface protein



100 YEARS AGO

Heredity — and variation too — are matters of which no naturalist likes to admit himself entirely careless. Everyone knows that, somewhere hidden among the phenomena denoted by these terms, there must be principles which, in ways untraced, are ordering the destinies of living things. Experiments in heredity have thus, as I am told, a universal fascination. All are willing to offer an outward deference to these studies. The limits of that homage, however, are soon reached, and, though all profess interest, few are impelled to make even the moderate mental effort needed to apprehend what has been already done... An eminent foreign professor lately told me that he believed there were not half a dozen in his country conversant with what may be called Mendelism, though he added hopefully, "I find these things interest my students more than my colleagues." A professed biologist cannot afford to ignore a new life-history, the Okapi, or the other last new version of the old story; but phenomena which put new interpretations on the whole, facts witnessed continually by all who are working in these fields, he may conveniently disregard as matters of opinion. Had a discovery comparable in magnitude with that of Mendel been announced in physics or chemistry, it would at once have been repeated and extended in every great scientific school throughout the world. William Bateson
From *Nature* 25 August 1904.

50 YEARS AGO

The theory of methods of crystal-structure determination is at an interesting stage... Mrs. D. Hodgkin, of Oxford, described a partial solution of the structure of vitamin B12; the molecule contains about a hundred atoms, almost three times as many as the most complicated molecule that has yet been dealt with successfully. Similar methods combined with the Fourier-transform concept have also enabled Sir Lawrence Bragg and M. F. Perutz, of Cambridge, to report some success with haemoglobin, the molecule of which contains about ten thousand atoms. Unfortunately, the Fourier projection obtained, having a resolution of only 4 Å., is not particularly clear, and some intense study will be required before any reliable deductions can be made from it.
From *Nature* 28 August 1954.

that is normally sensitive to protein-degrading enzymes (proteases) and has a turnover time of a few hours. In TSE-infected cells, by contrast, PrP is partly protease-resistant and quite stable (reviewed in ref. 10). Because it is the main component of purified infectious material⁴, this protein was itself proposed to be the infectious agent in TSEs⁹. Theoretically, the normal form of the protein, PrP^C (where 'C' denotes 'cellular'), converts rarely into an abnormal form (PrP^{Sc} for 'scrapie') that then catalyses the conversion of other normal molecules into the same abnormal form. The accumulated PrP^{Sc} then somehow causes neurodegeneration.

Proving the protein-only theory has been difficult, however. For example, because the infectious material is highly aggregated and heterogeneous in size, even the best preparations need 10⁵ PrP molecules to infect one mouse. So it has not yet been possible to prove that PrP is sufficient for infectivity by purifying the protein from infected brains.

Another approach has taken its cue from the 'amyloid' nature of PrP^{Sc}. This abnormal form — unlike PrP^C — contains a high percentage of β -sheets, one of the two main types of structural feature in proteins. Furthermore, PrP^{Sc} is filamentous and protease-resistant and shows birefringence on staining with the dye Congo Red. These four properties define amyloid, and gross amyloid deposits of PrP are seen in the brains of many CJD patients. Full-length, mature, soluble PrP^C has been converted to β -sheet-rich forms *in vitro* — but this material was not infectious.

There has been success on one front: the basic mechanism postulated for the TSE protein-only model has been demonstrated biochemically. PrP^{Sc} purified from infected brains can induce the conversion *in vitro* of small amounts of PrP^C to the PrP^{Sc} form, in a reaction that shows all the specificity of the TSE disease process¹¹. However, the limited extent of this reaction *in vitro* has so far made it impossible to demonstrate the generation of new infectivity, over that present in the input PrP^{Sc}.

How can one discover whether a protein is infectious? In budding yeast, clear evidence was provided by the genetic properties of the non-chromosomal genes [*URE3*] and [*PSI*⁺], which represent the prion forms of the Ure2 and Sup35 proteins, respectively¹². The same genetic approaches served to demonstrate that [*PIN*⁺] (also in budding yeast¹³) and [Het-s] (in the filamentous fungus *Podospora anserina*)¹⁴ represent the prion forms of the Rnq1 and HET-s proteins, respectively. But these methods are not feasible in mammals. However, [Het-s] can be efficiently transmitted to uninfected *P. anserina* cells by introducing amyloid fibrils of recombinant HET-s protein (not by introducing the soluble form or other aggregates of this protein). These results confirm

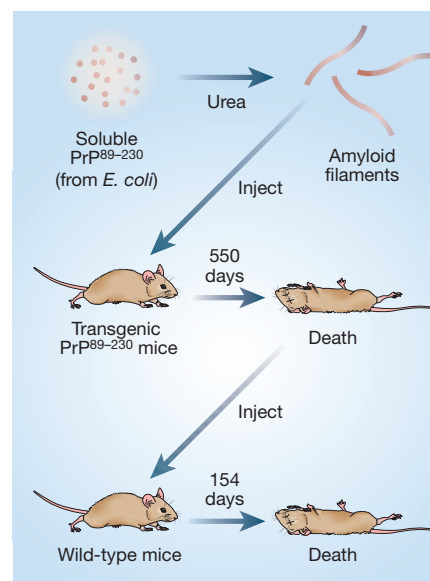


Figure 1 Can a protein be infectious? Legname *et al.*¹ engineered *Escherichia coli* bacteria to produce a fragment of PrP protein, consisting of amino acids 89–230. Adding urea to the soluble protein caused it to misfold and clump together into amyloid filaments. These were then injected into the brains of mice that had been engineered to express the same PrP fragment. About 550 days later, the animals died with a scrapie-like neurodegenerative disease. The authors also took brain tissue from these mice and injected it into wild-type animals; these animals, too, died from a scrapie-like disease. The findings argue that PrP alone is sufficient to cause transmissible spongiform encephalopathies.

that [Het-s] is a prion, and show that self-propagating amyloid is the infectious material¹⁵. Recently, similar results have been achieved in the [*PSI*⁺] system^{16,17}.

In the latest work, Legname *et al.*¹ engineered mice so that they lacked the normal PrP protein, instead producing a fragment of PrP (amino acids 89–230) at 16 times the usual levels. The idea was that this fragment might be more easily triggered than the full-length protein to convert to the abnormal form. The authors then prepared amyloid filaments of PrP⁸⁹⁻²³⁰ *in vitro* and inoculated the filaments into the animals' brains. They found that these mice died with a scrapie-like syndrome after about 550 days, whereas mice inoculated with a control buffer survived at least 620 days. Crucially, brain tissue from the amyloid-inoculated mice was then highly infectious when inoculated into normal mice (Fig. 1).

Unfortunately, the authors do not report whether the brains of the buffer-inoculated mice are also infectious. If they are not, then the results show that amyloid filaments composed of this PrP fragment can initiate an infectious process — apparently putting to rest the 'protein-only' dispute. If the brains of buffer-inoculated animals are infectious,

however, then one can conclude that overproduction of the PrP fragment results in prion formation *de novo* (perhaps accelerated by the injection of amyloid). The latter conclusion would also be a strong argument in favour of the protein-only model, as this is one of the genetic criteria for prions that were established by studies in yeast¹².

One other control would have been useful. The authors used amyloid filaments of the PrP fragment for infection, but they argue, paradoxically, that it is not amyloid that is the infectious form. For this reason, among others, injecting the soluble PrP fragment, or a non-specific aggregate, would have been helpful.

Nonetheless, this work adds to the already considerable evidence for the protein-only model. It is essential that the transgenic mice be made widely available, as this infectivity assay will be very useful in defining precisely the structure of the infectious material. ■

Herman K. Edskes and Reed B. Wickner are in the Laboratory of Biochemistry and Genetics, National Institutes of Health, 8 Center Drive, MSC 0830, Bethesda, Maryland 20892-0830, USA.
e-mail: wickner@helix.nih.gov

1. Legname, G. *et al.* *Science* **305**, 673–676 (2004).
2. Griffith, J. S. *Nature* **215**, 1043–1044 (1967).
3. Dickinson, A. G., Meikle, V. M. H. & Fraser, H. J. *Comp. Pathol.* **78**, 293–299 (1968).
4. Bolton, D. C., McKinley, M. P. & Prusiner, S. B. *Science* **218**, 1309–1311 (1982).
5. Oesch, B. *et al.* *Cell* **40**, 735–746 (1985).
6. Chesebro, B. *et al.* *Nature* **315**, 331–333 (1985).
7. Carlson, G. A. *et al.* *Cell* **46**, 503–511 (1986).
8. Bueler, H. *et al.* *Cell* **73**, 1339–1347 (1993).
9. Prusiner, S. B. *Science* **216**, 136–144 (1982).
10. Harris, D. A. *Br. Med. Bull.* **66**, 71–85 (2003).
11. Kocisko, D. A. *et al.* *Nature* **370**, 471–474 (1994).
12. Wickner, R. B. *Science* **264**, 566–569 (1994).
13. Derkatch, I. L. *et al.* *Cell* **106**, 171–182 (2001).
14. Coustou, V. *et al.* *Proc. Natl Acad. Sci. USA* **94**, 9773–9778 (1997).
15. Maddelein, M. L. *et al.* *Proc. Natl Acad. Sci. USA* **99**, 7402–7407 (2002).
16. King, C. Y. & Diaz-Avalos, R. *Nature* **428**, 319–323 (2004).
17. Tanaka, M. *et al.* *Nature* **428**, 323–328 (2004).

Galaxy formation

Caught in the act?

Zoltán Haiman

Which came first, the stars and gas that make up a galaxy, or the giant black hole at its centre? Observations of a distant galaxy, caught as it forms, could help solve this chicken-and-egg problem.

Galaxies are thought to be surrounded by massive haloes of dark matter, each outweighing its galaxy by a factor of about eight. The visible part of a galaxy, occupying the inner 10% of the halo, consists of a mixture of stars and gas. Galaxies harbour a giant black hole at their centres, which in some cases is actively fuelled as it sucks in surrounding gas. In especially active galaxies, called quasars, the fuelling rate is so high that the radiation generated close to the black hole outshines the cumulative starlight from the entire galaxy. The sequence of cosmic events that leads to this configuration is still largely mysterious. How does gas condense into the central regions of the dark-matter halo? At what stage of the gas condensation process do the stars and the giant black hole light up? The unique observation of a distant quasar by Weidinger *et al.*¹, reported on page 999 of this issue, offers fresh insight.

The formation of massive dark-matter haloes is dictated by gravity, and can be described by using *ab initio* calculations². As the Universe expanded from its dense beginning, tiny inhomogeneities in the distribution of dark matter were amplified through the effects of gravity. Regions of space that were slightly denser than average had a higher gravitational pull on their surroundings; eventually, these regions stopped following

the expansion of the rest of the Universe, turned around and re-collapsed on themselves. The resulting dense knots of dark matter — forming the intersections of a cosmic web of less-dense dark-matter filaments — are believed to be the sites at which galaxies lit up.

Dark matter thus dominates the formation of a galaxy, at least initially, and determines the gross properties of the galaxy population, such as their abundance, size and spatial distribution. But it is the trace amount of gas (mostly hydrogen and helium), pulled with the dark matter into the collapsed haloes, that forms the visible parts of galaxies and determines their observable properties. In particular, to condense to the core of the dark halo, the gas must cool continuously so as to deflate the pressure acquired by its compression. A fraction of the gas (typically 10% by mass) eventually turns into stars, and a much smaller fraction (typically 0.1%) into the central massive black hole³.

The composition of the gas inside the galaxy can be studied through the spectrum of radiation that it absorbs and emits. Primordial gas is essentially a pure mix of hydrogen and helium, but the spectra of all of the quasars discovered so far have shown the presence of various heavier elements (such as carbon, nitrogen, oxygen and iron). This

indicates that the gas has been enriched by the nucleosynthetic yields from previous generations of stars. Even the most distant quasars, including those that existed about a billion years after the Big Bang (a mere 5% of the current age of the Universe), show a significant heavy-element content⁴. This suggests that vigorous star-formation is a necessary condition for any quasar activity. On the other hand, star formation seems to be occurring on relatively small scales, close to the galactic centre. A natural inference would then be the following sequence of events: the cosmic gas first contracts to the inner regions of the halo, and only then forms stars — but this is still before the formation (or at least activation) of any central quasar black hole.

Not necessarily so, according to Weidinger *et al.*¹. They have detected the faint glow of hydrogen emission enveloping a distant quasar at a radius equivalent to about 100,000 light years — several times the size of the visible part of a typical galaxy. Such emission has a simple physical origin. The hydrogen atoms falling through the halo are ionized by the quasar's light, then recombine with electrons to become atoms again. Each recombination results in the emission of a so-called Lyman- α photon (a photon with energy equal to the difference between the ground and first excited states of a hydrogen atom). As a result, when viewed through a filter tuned to the Lyman- α frequency, a faint 'fuzz' can be seen to surround quasars⁵. This fuzz can serve as a diagnostic of whether or not a spatially extended distribution of infalling gas is present around the quasar⁶. If most of the gas has already cooled and settled at the centre of the halo, the extended fuzz would be absent.

Although the fuzz detected by Weidinger *et al.* is faint, it is as bright as would be expected if all of the hydrogen needed to make up a typical galaxy is still infalling⁶. As a result, the galaxy imaged by Weidinger *et al.* is likely to be still in its infancy, despite the fully formed appearance of its bright central quasar black hole. From the shape and kinematics of the fuzz, the authors were also able to confirm the presence of the dark-matter halo, which is accelerating the infall of the hydrogen gas, and to measure the halo's mass. Their value — $2\text{--}7 \times 10^{12}$ solar masses — is in accord with independent estimates from spectral absorption features⁷ and from the abundance⁸ of other quasars of similar brightness.

Extended Lyman- α emission has previously been detected around quasars, but its interpretation as being due to a wide region of infalling gas was less convincing^{9–11}. In contrast, Weidinger *et al.*¹ have obtained a two-dimensional image of the source that clearly shows the extent of the emission, and a spectrum showing a velocity pattern consistent with infall. Extended emission is a relatively common phenomenon among

Plant biology

The benefits of nicotine

Gardeners and farmers know that nicotine has its uses: a quick squirt of a nicotine solution keeps pests at bay. But do the plants that make this compound use it to deter invaders? This isn't as obvious as one might think; evolution might have rendered the insects that eat these particular plants completely resistant to nicotine. Anke Steppuhn *et al.* aimed to find out (*PLoS Biol.* **2**, e217; 2004).

They produced transgenic

tobacco plants (*Nicotiana attenuata*) in which a key enzyme in nicotine synthesis was silenced. The nicotine concentration in these plants dropped to 3–4% of normal levels. Experiments with greenhouse-grown tobacco showed that the larvae of a key pest — the tobacco hornworm *Manduca sexta* (pictured) — much preferred leaves from the transgenic plants than from wild-type plants. The larvae also grew much more quickly when reared on

the transgenic plants, implying that although these bugs can tolerate nicotine, they fare better without it.

Experiments with field-grown tobacco gave similar results, and showed that the transgenic plants also lost significantly more leaf area to pests than did wild-type plants. The inescapable conclusion is that secondary metabolites such as

nicotine, although not essential for normal plant growth and reproduction, nonetheless make significant contributions to ecology.

Amanda Tromans



E. HABEGGER/GRANT HEILMAN PHOTOGRAPHY

so-called radio-loud quasars — a special subset believed to drive powerful gaseous outflows. However, the gas surrounding these sources is always enriched with heavy elements, whereas the fuzz detected by Weidinger *et al.* appears to be pristine (no heavy-element absorption or emission features are detected).

This result will undoubtedly prompt further theoretical modelling of the density and velocity distribution of the infalling gas. One outstanding question, for example, is whether the gas could have been delivered by a recent merger with another galaxy; such mergers stir up and spread gas over large regions, and might also be responsible for having activated the quasar in the first place. An observational census of similar quasars should be revealing, and feasible with a modest investment of time on modern telescopes. It could show whether prompt activation of a quasar, at an early stage in the assembly of the galaxy, while there is still a significant amount of cosmic hydrogen-gas infall, is the norm among galaxies, or whether it is a peculiarity of the galaxy examined by Weidinger and colleagues. In the meantime, we have a new piece in the chicken-and-egg puzzle, suggesting that, at least occasionally, the egg — the black hole — may come first.

Zoltán Haiman is in the Department of Astronomy, Columbia University, New York, New York 10027, USA.

e-mail: zoltan@astro.columbia.edu

- Weidinger, M., Möller, P. & Fynbo, J. P. U. *Nature* **430**, 999–1001 (2004).
- Navarro, J. E., Frenk, C. S. & White, S. D. M. *Astrophys. J.* **462**, 563–575 (1996).
- Magorrian, J. *et al.* *Astron. J.* **115**, 2285–2305 (1998).
- Fan, X. *et al.* *Astron. J.* (in the press); preprint at <http://arxiv.org/abs/astro-ph/0405138> (2004).
- Rees, M. J. *Mon. Not. R. Astron. Soc.* **231**, 91–95 (1988).
- Haiman, Z. & Rees, M. J. *Astrophys. J.* **556**, 87–92 (2001).
- Barkana, R. & Loeb, A. *Nature* **421**, 341–343 (2003).
- Haiman, Z. & Hui, L. *Astrophys. J.* **547**, 27–38 (2001).
- Möller, P., Warren, S. J., Fall, S. M., Jakobsen, P. & Fynbo, J. U. *ESO Messenger* **99**, 33–35 (2000).
- Bunker, A., Smith, J., Spinrad, H., Stern, D. & Warren, S. *Astrophys. Space Sci.* **284**, 357–360 (2003).
- Bergeron, J. *et al.* *Astron. Astrophys.* **343**, L40–L44 (1999).

Immunology

Aid for AID

Almudena R. Ramiro and Michel C. Nussenzweig

Activation-induced deaminase catalyses two processes that diversify antibodies. But this enzyme need not work alone: a partner links it to its substrate — single-stranded DNA — and to DNA-repair molecules.

One of the cardinal features of immune systems is that they show enough diversity to mount a defence against virtually any pathogen. The B cells of the immune system, for instance, produce a staggering repertoire of protective antibody proteins. In human and mouse B cells, three DNA-modifying processes generate such diversity: V(D)J recombination, somatic hypermutation and class switch recombination. The discovery that a single enzyme — activation-induced deaminase (AID) — is both necessary^{1,2} and sufficient for the latter two processes was a breakthrough in understanding the underlying molecular mechanisms. On page 992 of this issue, Chaudhuri and colleagues³ advance our knowledge further, revealing a protein that can assist in AID-induced antibody diversification.

Each of us inherits a limited set of antibody, or immunoglobulin, genes. But the number of pathogens that we might encounter is potentially limitless — hence the requirement for DNA-modifying processes to generate antibody diversity. An antibody contains one of several specific constant regions (which spark off different immune responses), joined to a variable region (which recognizes a particular foreign molecule, or antigen). V(D)J recombination, which occurs in immature B cells, generates the variable region, by randomly shuffling the V, D and J segments of the immunoglobulin genes (Fig. 1). This produces a spread of B cells whose antibodies can recognize different antigens.

Two other processes further diversify antibody genes in mature B cells that have encountered their cognate antigen. As the name suggests, class switch recombination changes the class of antibody that is produced so as to elicit the most appropriate immune response. This occurs by combining the variable region of the gene with appropriate constant regions. Meanwhile, somatic hypermutation introduces single mutations to the variable region, potentially increasing the antibody's affinity for its antigen.

Several years ago, it was discovered that both class switch recombination and somatic hypermutation require the AID protein^{1,2}. But it has been less easy to work out exactly what AID does. The protein shows a close relationship to APOBEC-1, an enzyme that converts cytosine nucleotides to uracils in RNA. So it was initially suggested¹ that AID also 'edits' RNA — specifically, a messenger RNA that encodes an endonuclease. Endonucleases are enzymes that generate nicks in DNA; the thinking was that the endonuclease in question, appropriately edited, would thereby initiate class switch recombination and somatic hypermutation.

An alternative idea was that AID might deaminate cytosine nucleotides in DNA, not RNA^{1,4}. In this model, the precise details of subsequent events would depend on how cells attempt to repair the cytosine-to-uracil mutations (uracils in DNA are not usually tolerated). Although there is no definitive proof for this theory, it is supported by several observations. First, AID mutates cytosines to uracils in *Escherichia coli*³. These bacteria

should not express the putative target mRNA, arguing against the RNA-editing model. Second, AID can bind to and deaminate DNA (in single-stranded form)^{5–9}. Third, AID associates with the immunoglobulin-gene region in B cells undergoing class switch recombination¹⁰. And finally, uracil-*N*-glycosylase, an enzyme that removes uracils from DNA, is required for normal somatic hypermutation and class switch recombination^{11,12}. Chaudhuri and colleagues' latest findings³ lend additional strong support to the DNA-deamination idea.

DNA is generally double-stranded. But, as mentioned above, AID targets single-stranded DNA. How could such a single-stranded substrate arise *in vivo*? The answer seems to lie in the fact that single-stranded regions of DNA are transiently exposed on the surface of the enzyme RNA polymerase during transcription — the process by which genes are copied into mRNA. These exposed regions are generally on the opposing strand to that bearing the gene being expressed. So, in *E. coli*, AID targets the single-stranded region that is complementary to a transcribed gene⁹. Transcription of appropriate regions of the immunoglobulin genes also seems to be a prerequisite for somatic hypermutation and class switch recombination.

But that's not the end of the story. In B cells undergoing somatic hypermutation, mutations can be made in either DNA strand. And there are certain immunoglobulin-gene hypermutation 'hotspots', defined by the sequence RGYW (where R is adenine or guanine, G is guanine, Y is cytosine or thymine, and W is adenine or thymine)¹³. This sequence preference is due at least in part to AID itself, as the deamination of single-stranded DNA by AID *in vitro* is skewed to RGYW motifs⁸.

But how is AID targeted to these single-stranded regions during somatic hypermutation? Chaudhuri *et al.* previously⁵ made the key observation that AID purified from B cells deaminated DNA that was provided in single-strand conformation, but was inactive on DNA substrates — akin to the variable regions of immunoglobulin genes — that had been transcribed by the T7 RNA polymerase enzyme. In contrast, crude AID preparations were active in both reactions. The implication was that AID does not function alone to target and deaminate these substrates.

Chaudhuri *et al.*³ have now used a classical biochemical approach to search for factors that might work with AID. They found replication protein A (RPA), a protein that binds single-stranded DNA and is important in various aspects of DNA metabolism. The authors discovered that transcription-dependent deamination by AID *in vitro* depends on RPA, and that it correlates with the RGYW content of the DNA substrate. Furthermore, a physical interaction between these two proteins is required for

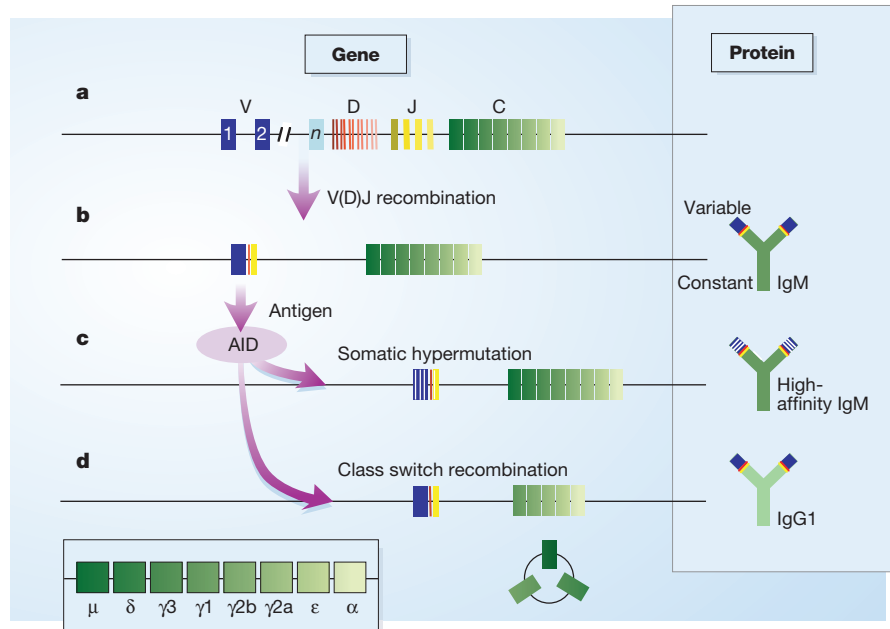


Figure 1 Generating antibody diversity. Antibodies are encoded by immunoglobulin genes; these include V, D and J regions (which produce the variable segment of an antibody protein) and C regions (which produce the constant portion). a, In an immature B cell, an immunoglobulin gene includes the full range of V, D, J and C regions. b, V(D)J recombination shuffles and excises some of the V, D and J regions, generating an antibody (IgM) with a variable region that recognizes a particular antigen. When the B cell meets this antigen, two other processes — both catalysed by the AID protein — are triggered. c, Somatic hypermutation generates mutations (thin white lines) in the variable regions, potentially generating an IgM with higher affinity for its antigen. d, Class switch recombination results in the excision of some of the C regions, generating a different class of antibody (IgG1 is shown here). Chaudhuri *et al.*³ have found that AID can work with a partner to achieve somatic hypermutation.

AID to bind to DNA, in an RGYW-dependent manner, during T7-polymerase-mediated transcription. These findings suggest that RPA stabilizes the single-stranded DNA produced during transcription, to enhance AID activity. It is not difficult to imagine that RPA could also serve as an entry point for additional mutational factors and for DNA-repair factors — such as uracil-*N*-glycosylase and mismatch-repair proteins — that are known to associate with RPA.

Chaudhuri *et al.* also show that only phosphorylated AID — which is produced specifically by B cells — interacts with RPA. Because the overexpression of AID in non-B cells is sufficient to support RGYW-focused somatic hypermutation and class switch recombination, it is generally accepted that AID is the only B-cell-specific factor that is required for these processes. However, in non-B cells mutations are strongly biased towards guanines and cytosines, whereas in B cells, mutations are found on all four nucleotides. Chaudhuri and colleagues' findings suggest that a B-cell-specific interaction between RPA and phosphorylated AID is required to support bona fide somatic hypermutation, and may explain the guanine:cytosine bias in cells in which such an interaction cannot occur.

Chaudhuri and colleagues' findings³ enrich our understanding of somatic hypermutation, by defining AID's first partner and by providing a conceptual link between AID,

single-stranded DNA exposed during transcription, and DNA repair. Their exciting results also raise several questions. How does RPA contribute *in vivo*? How is the phosphorylation of AID regulated? If the RPA–AID complex specifically targets hypermutation hotspots, how does hotspot targeting occur in non-B cells? Are uracil-*N*-glycosylase or mismatch-repair proteins recruited to the deamination site through RPA, through AID, or through the AID–RPA complex? Finding this complex is an essential first step towards answering such questions. It also, coincidentally, lends weight to the idea that AID works on DNA, not RNA.

Almudena R. Ramiro and Michel C. Nussenzweig are at the Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.

e-mail: nussen@mail.rockefeller.edu

1. Muramatsu, M. *et al.* *Cell* **102**, 553–563 (2000).
2. Revy, P. *et al.* *Cell* **102**, 565–575 (2000).
3. Chaudhuri, J., Khuong, C. & Alt, F. W. *Nature* **430**, 992–998 (2004).
4. Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. *Nature* **418**, 99–103 (2002).
5. Chaudhuri, J. *et al.* *Nature* **422**, 726–730 (2003).
6. Bransteitter, R., Pham, P., Scharff, M. D. & Goodman, M. F. *Proc. Natl Acad. Sci. USA* **100**, 4102–4107 (2003).
7. Dickerson, S. K., Market, E., Besmer, E. & Papavasiliou, F. N. *J. Exp. Med.* **197**, 1291–1296 (2003).
8. Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. *Nature* **424**, 103–107 (2003).
9. Ramiro, A. R. *et al.* *Nature Immunol.* **4**, 452–456 (2003).
10. Nambu, Y. *et al.* *Science* **302**, 2137–2140 (2003).
11. Di Noia, J. & Neuberger, M. S. *Nature* **419**, 43–48 (2002).
12. Imai, K. *et al.* *Nature Immunol.* **4**, 1023–1028 (2003).
13. Storb, U. & Stavnes, J. *Curr. Biol.* **12**, R725–R727 (2002).

Chemistry

Attractive development for male moths

Green Chem. **6**, 305–307 (2004)

The larva of the cabbage moth (*Mamestra brassicae*) is a voracious creature that feeds on a wide variety of vegetable leaves and is a serious pest for farmers in North America and Europe. Sex pheromones are used by the ton to control these and numerous other insect pests by luring male moths into traps at the edges of fields, thus stopping them from breeding.

Petra Nešňerová *et al.* now report a greatly simplified synthesis of a moth sex pheromone that may also be applicable to other lepidopteran pheromones. The authors genetically modified tobacco plants (*Nicotiana tabacum*) so that they produced a molecular precursor to the pheromone, bypassing most of the chemical steps required during conventional synthesis. They claim that this is the first synthesis of a sex pheromone that exploits such 'molecular farming' techniques.

The pheromone precursor, an ester, made up 6.5% of the total lipids extracted from the plant. Two simple chemical steps then delivered the final pheromone, which performed well in tests in the cabbage fields of Bohemia.

The authors are now working on a genetic-modification strategy that could produce the pheromones directly in the tobacco plants.

Mark Peplow

Molecular biology

Codons and gene expression

Proc. Natl Acad. Sci. USA **101**, 12588–12591 (2004)

In the genetic code, three bases make up a codon, and a codon instructs the cellular machinery to insert a particular amino acid into a protein. Most amino acids can be encoded by more than one codon, and Joshua B. Plotkin *et al.* have found hints that different human tissues preferentially use different types of codon.

The researchers analysed the sequences of genes that are selectively expressed in the brain, liver, uterus, testis, ovary or vulva. In most cases, they found that each tissue uses a characteristic set of codons for a particular amino acid. For example, a testis-specific gene uses the codon GCC for the amino acid alanine only 20% of the time, whereas a uterus-specific gene uses the same codon nearly 60% of the time.

The authors suggest that the differing patterns of codon usage may help determine which genes will be efficiently converted into proteins in different tissues. This could be the case if, for example, the different tissues contain different pools of transfer RNAs — the molecules that recognize the codon and

slot appropriate amino acids into the protein. The finding might be important to consider during gene therapy, so that a gene inserted into a target tissue can be translated as efficiently as possible.

Helen Pearson

Biological chemistry

Bromeliad prospects

J. Natural Products doi: 10.1021/np0499766 (2004)

Bromeliads are common tropical plants that hold water in a pool, or tank (pictured), formed by the bases of their leaves. These tanks are home to a rich assortment of organisms, including many as yet uncultured bacteria that produce a host of unknown, but possibly useful, chemicals. Sean F. Brady and Jon Clardy now report the identification of an antibiotic produced by such a bacterium.

Brady and Clardy extracted DNA fragments from the tank water of Costa Rican bromeliads, and inserted them into



the genome of *Escherichia coli* bacteria. One of the *E. coli* clones produced a compound with antibiotic activity, which the authors identify as palmitoylputrescine: they infer that this small, organic molecule is synthesized through a gene encoding an enzyme called palmitoylputrescine synthase.

The function of palmitoylputrescine in bromeliad bacteria is unclear, but similar molecules are involved in activating multiple receptors in animal tissues and inducing volatile defence signals in plants. The researchers hope that, with further refinement, the method they have used will help in testing DNA extracted from other environmental samples for genes that encode potentially valuable natural products.

Helen Pilcher

Optics

Light takes a turn

Phys. Rev. Lett. **93**, 083901 (2004)

Electrons moving perpendicular to a magnetic field experience a force that pushes them off course. In a two-dimensional conductor they tend to pile up on one side, creating a potential

difference. This phenomenon, called the Hall effect, is used to detect magnetic fields by means of the Hall voltage they generate in a suitable sensor. Masaru Onoda *et al.* show theoretically that something like the Hall effect operates for photons moving in a ray of light when they travel through a medium with a non-uniform refractive index.

For example, circularly polarized light passing across the interface between two materials with different refractive indices is shifted perpendicular to both the incident direction of the beam and the normal vector of the interface. Where does the force come from in this case? It is a consequence of the polarization, or spin, of a photon. The spin and angular momentum of a photon change as it encounters a change in refractive index, and in order to conserve total momentum this leads to a shift in the phase of the electromagnetic wave — it is a purely geometric effect.

The phase shift is known as the Berry phase, and it means that the trajectory of the photon alters. Onoda *et al.* say that the effect might be magnified, so as to be easily observed, in photonic crystals (regular, microscopic arrays of light-scatterers) with non-uniform spacings between the scatterers.

Phillip Ball

Aquaculture

Signs of escaped salmon

Ecol. Freshwater Fish **13**, 176–184 (2004)

Escapes of Atlantic salmon (*Salmo salar*) from fish farms are a concern for ecologists, who fear that they may interbreed with wild populations and reduce their genetic diversity. J. B. Dempson and M. Power now propose a non-lethal way to identify escapees by spotting the tell-tale signs of the diet they would have eaten while in captivity.

Dempson and Power took wild salmon from the Conne River in Newfoundland, Canada, and salmon from farms in the Bay d'Espoir at the river's mouth. They sampled adipose fin and muscle, and measured the concentrations of the stable isotopes carbon-13 and nitrogen-15 in these two tissues, after adjusting for fat content.

Tissues from farmed salmon contained 0.126% less carbon-13 than those from wild salmon, whereas levels of nitrogen-15 were 0.179% higher on average. There was no clear difference in isotopic signature between adipose fin and muscle. The isotopic contrast between the farmed and the wild fish is due to their differing diets. Although it is unclear how long escaped salmon would retain their characteristic signature in the wild, the authors believe that it should remain evident for an entire growth season.

Michael Hopkin

Event timing turns punishment to reward

Linking a smell with an electric shock does not always have an aversive effect in flies.

Can a relief from pain be a pleasure? If so, noxious events should — despite their typically aversive effects — also have a ‘rewarding’ after-effect^{1–3}. Through training fruitflies by using an electric shock paired with an odour, we show here that the shock can condition either avoidance of this odour or approach to it. These opposing behaviours depend on the relative timing of the shock and odour presentations during training, and indicate that a shock can act as either an aversive reinforcer or an appetitive one.

To measure both aspects of these bidirectional behavioural responses within the same set-up, we used fruitflies (*Drosophila melanogaster*) that had undergone odour-discrimination learning reinforced by electric shock⁴ (Fig. 1a). All experimental groups received the same amount of odour–shock training. The only variable was the interstimulus interval (ISI), which was the interval between the onset times

of exposure to the odour for association (the ‘trained’ odour) and to the shock (Fig. 1a). Training sessions were repeated four times and were separated by a 20-minute rest in a food vial.

The conditioned behaviour was tested 15 min after training in a forced-choice situation, by counting how many animals chose either a control or the trained odour (odour A or B, respectively; Fig. 1a). Positive learning indices indicate conditioned avoidance of the trained odour, whereas negative scores indicate conditioned approach to it.

During testing, flies showed opposite responses to the trained odour (either conditioned avoidance or approach), depending on the temporal sequence of odour and shock that they had experienced during training (Fig. 1b). If the odour preceded the shock, flies showed conditioned avoidance (for ISIs of -23 s and -3 s; $P < 0.005$; Fig. 1b). However, when the shock preceded the odour, flies showed conditioned approach (for ISIs of $+32$ s and $+42$ s; $P < 0.005$; Fig. 1b). We conclude that flies were able to associate the same odour with either danger or safety. Control groups trained using very long ISIs gave no evidence of learning (for example, for ISIs of -83 s or $+187$ s; $P > 0.005$ for forward or backward control, respectively; Fig. 1b).

We found that the effect of shock turns

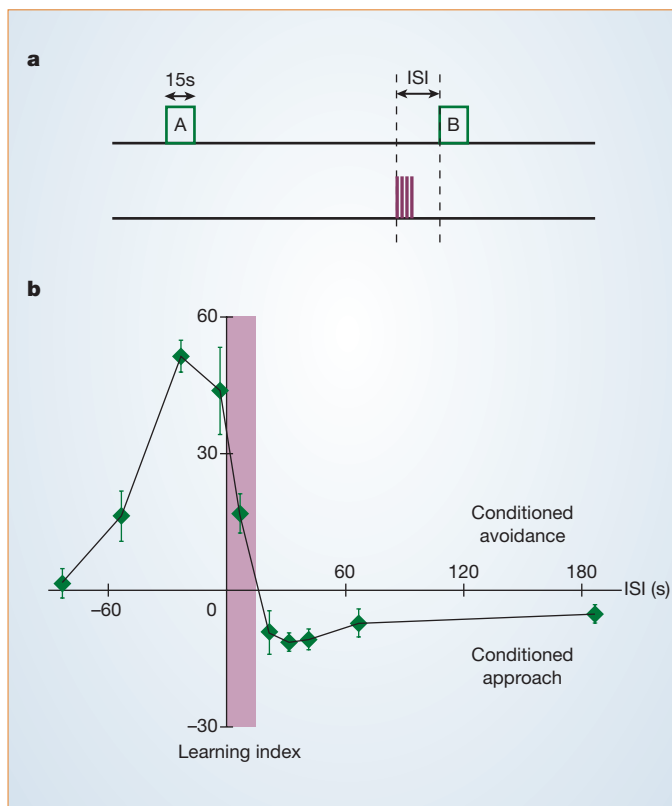


Figure 1 Bidirectional behavioural plasticity in associative learning. **a**, Experimental procedure. Odour stimuli for conditioning: benzaldehyde and 3-octanol (ref. 11; presented for 15 s each; green boxes); unconditioned stimulus: electric shock (90 volts, 4 pulses within 16 s; pink lines). The odours are aversive to naive flies. The interstimulus interval (ISI) between the onset of the trained odour B and the shock is varied between groups, whereas the interval between the control odour A and shock is the same for all groups (-150 s). Relative preferences for odours A and B were compared between reciprocally trained groups (that is, the ones for which benzaldehyde was the trained odour were compared with the ones for which 3-octanol was the trained odour); the learning index was calculated as the difference between these preferences divided by two, to yield values between 100 and -100 . **b**, Systematic analysis of conditioned behaviour (learning index, mean $\pm$ s.e.m.) as a function of the ISI. Pink shading indicates duration of the electric shock. Conditioned behaviour was significant for the groups with ISI of -23 s, -3 s, $+32$ s and $+42$ s ($P < 0.005$; one-sample *t*-test against zero). The significance level of each test was adjusted to 0.005 with a Bonferroni correction to avoid false positives by multiple tests.

from punishing to rewarding in a time window around shock application (pink shading in Fig. 1b). This indicates that odours can act as predictors of danger when they precede shock during training but, owing to a long-lasting after-effect of shock, they can also be used to predict safety when they follow shock during training. Conditioned avoidance was stronger than conditioned approach ($P < 0.05$, *t*-test: ISI, -23 s compared with $+32$ s). This quantitative comparison is possible because the two aspects of timing-dependent behavioural plasticity were directly measured within the same set-up, rather than indirectly^{2,5}.

Bidirectional synaptic plasticity has a comparable dependence on timing: the sequence of two inputs determines whether synapses are potentiated or depressed^{6–8}. This characteristic would lead to bidirectional associative learning if it occurred during association formation at the neuronal convergence site of odour and shock. Alternatively, the dual and opposing behavioural effects of shock could reflect a bidirectional modulation of internal reinforcement signalling, as found in mammalian dopaminergic neurons^{9,10}. It will be interesting to investigate whether the appetitive effect of shock in flies shares a common neuronal circuitry with reward processing⁴. The detailed characterization of the rewarding after-effect of negative reinforcement should advance

our understanding of the behavioural consequences of traumatic experience.

Hiromu Tanimoto, Martin Heisenberg, Bertram Gerber

Lehrstuhl für Genetik und Neurobiologie, Biozentrum, Universität Würzburg, Am Hubland, 97074 Würzburg, Germany
e-mail: hiromut@biozentrum.uni-wuerzburg.de

- Solomon, R. L. & Corbit, J. D. *Psychol. Rev.* **81**, 119–145 (1974).
- Rescorla, R. A. & LoLordo, V. M. *J. Comp. Physiol. Psychol.* **59**, 406–412 (1965).
- Wagner, A. R. in *Information Processing in Animals: Memory Mechanisms* (eds Spear, N. E. & Miller, R. R.) 5–47 (Erlbaum, Hillsdale, New Jersey, 1981).
- Schwaerzel, M. et al. *J. Neurosci.* **23**, 10495–10502 (2003).
- Hellstern, E., Malaka, R. & Hammer, M. *Learn. Mem.* **4**, 429–444 (1998).
- Bi, G.-Q. & Poo, M.-M. *Annu. Rev. Neurosci.* **24**, 139–166 (2001).
- Frome, R. C. & Dan, Y. *Nature* **416**, 433–438 (2002).
- Abbott, L. F. & Nelson, S. B. *Nature Neurosci.* **3**, 1178–1182 (2000).
- Tobler, P., Dickinson, A. & Schultz, W. *J. Neurosci.* **23**, 10402–10410 (2003).
- Ungless, M. A., Magill, P. J. & Bolam, J. P. *Science* **303**, 2040–2042 (2004).
- Schwaerzel, M., Heisenberg, M. & Zars, T. *Neuron* **35**, 951–960 (2002).

Competing financial interests: declared none.

brieF communications arising online
www.nature.com/bca

Mitochondrial permeability: Dual role for the ADP/ATP translocator?

A. P. Halestrap (doi:10.1038/nature02816)

Mitochondrial permeability

Dual role for the ADP/ATP translocator?

Arising from: J. E. Kokoszka *et al.* *Nature* **427**, 461–464 (2004)

The ADP/ATP translocator (or adenine nucleotide translocase; ANT) is thought to play a dual role: in the transport of ADP and ATP across the mitochondrial inner membrane and in the formation of the mitochondrial permeability-transition pore (mtPTP), a nonspecific pore that is an important mediator of apoptosis (programmed cell death)^{1–3}. However, Kokoszka *et al.*⁴ have shown that mitochondria from livers of ‘ANT-knockout’ mice, in which the ANT has been genetically inactivated, still possess mtPTP activity. From this, the authors conclude that the ANT is a non-essential component of the mtPTP that may be dispensable for mtPTP-associated cell death⁴. These results, which contradict previous evidence^{1,2} and cast doubt on a widely accepted model for the mtPTP (ref. 1), warrant scrutiny and call for a fundamental reappraisal of the role of the ANT in liver metabolism.

Nearly all essential metabolic pathways in the liver require an abundant supply of cytosolic ATP, with gluconeogenesis, urea synthesis and lipogenesis being the most demanding. Most of this ATP is supplied by mitochondrial oxidative phosphorylation and is exported to the cytosol by the ANT in exchange for incoming ADP—as illustrated, for example, by the inhibition of hepatic gluconeogenesis and urea synthesis by atracyloside (a specific inhibitor of the ANT)⁵. It is therefore surprising that mice without hepatic ANT can survive the loss of these essential metabolic pathways, let alone live normally for at least a year⁴.

What metabolic changes could account for this unexpected finding? The liver might switch to using much more glucose in order to generate ATP by glycolysis in the cytosol, but this would lead to massive production of lactic acid by the liver (and subsequent oxidation by red muscle and heart) that should be readily detected as an increase in blood lactate concentrations. And even this compensatory mechanism would not enable the liver to carry out gluconeogenesis, a task that would have to be taken over by the kidney. An alternative is that another member of the mitochondrial carrier family, such as the deoxynucleotide carrier or the ATP–Mg/P_i exchange carrier, might take over the role of the ANT³.

Or could it be that the mitochondria are not totally devoid of ANT activity, as Kokoszka *et al.* claim? Their indirect assay for this activity is not particularly sensitive, relying as it does on stimulation of the already enhanced rate of state-3 respiration, although any residual activity must be relatively small.

The principal conclusion reached by the authors⁴, that the ANT is not an essential component of the mtPTP, may also require qualification in the light of evidence that the ANT, together with cyclophilin D, does fulfil this role^{1,2}. Putting aside the possibility that there may be some residual ANT present, could there be an alternative explanation for the new observations?

The ANT is the most abundant member of a family of transporters that have a conserved structure³, and several members of this family will form channels under specific conditions^{1,6,7}; it would be consistent with the

authors’ data if the ANT represented the normal target for cyclophilin-D binding that leads to mtPTP formation. However, in the absence of the ANT, another less abundant member of the carrier family might take its place, albeit, as observed by the authors, with less sensitivity to calcium and no sensitivity to ligands of the ANT.

Also, it has been proposed⁸ that the mtPTP may be formed from aggregates of denatured membrane proteins: these may normally be stabilized by a chaperone, but open into a pore when acted upon by cyclophilin D in the presence of calcium.

There is strong evidence that the mtPTP can be formed from purified ANT and cyclophilin D^{2,7,9}, but this falls short of proof. More rigorous analysis of the reconstitution of the mtPTP from purified components should resolve this controversy. Such an analysis will be facilitated by the X-ray crystallographic structures of the ANT (ref. 10) and cyclophilin D (ref. 11).

Andrew P. Halestrap

Department of Biochemistry, University of Bristol, Bristol BS8 1TD, UK

e-mail: a.halestrap@bristol.ac.uk

doi:10.1038/nature02816

1. Halestrap, A. P. & Brenner, C. *Curr. Medicinal Chem.* **10**, 1507–1525 (2003).
2. Halestrap, A. P., Doran, E., Gillespie, J. P. & O’Toole, A. *Biochem. Soc. Trans.* **28**, 170–177 (2000).
3. Palmieri, F. *Pflügers Arch.* **447**, 689–709 (2004).
4. Kokoszka, J. E. *et al.* *Nature* **427**, 461–464 (2004).
5. Stubbs, M., Vignais, P. V. & Krebs, H. A. *Biochem. J.* **172**, 333–342 (1978).
6. Dierks, T., Salentin, A. & Krämer, R. *Biochim. Biophys. Acta* **1028**, 281–288 (1990).
7. Brustovetsky, N., Tropschug, M., Heimpe, S., Heidkamper, D. & Klingenberg, M. *Biochemistry* **41**, 11804–11811 (2002).
8. He, L. H. & Lemasters, J. J. *FEBS Lett.* **512**, 1–7 (2002).
9. Crompton, M., Virji, S. & Ward, J. M. *Eur. J. Biochem.* **258**, 729–735 (1998).
10. Pebay-Peyroula, E. *et al.* *Nature* **426**, 39–44 (2003).
11. Waldmeier, P. C., Zimmermann, K., Qian, T., Tintelnot-Blomley, M. & Lemasters, J. J. *Curr. Medicinal Chem.* **10**, 1485–1506 (2003).

The synthesis of organic and inorganic compounds in evolved stars

Sun Kwok

Institute of Astronomy & Astrophysics, Academia Sinica, PO Box 23-141, Taipei 106, Taiwan

Recent isotopic analysis of meteorites and interplanetary dust has identified solid-state materials of pre-solar origin. We can now trace the origin of these inorganic grains to the circumstellar envelopes of evolved stars. Moreover, organic (aromatic and aliphatic) compounds have been detected in proto-planetary nebulae and planetary nebulae, which are the descendants of carbon stars. This implies that molecular synthesis is actively happening in the circumstellar environment on timescales as short as several hundred years. The detection of stellar grains in the Solar System suggests that they can survive their journey through the interstellar medium and that they are a major contributor of interstellar grains.

Work on stellar nucleosynthesis in the 1950s has led to our current realization that most of the chemical elements are synthesized in stars. Helium is made by hydrogen burning in the core during the main sequence and in a shell above the core in the red giant phase. The element carbon is created by helium-burning (the triple- α process), first through core burning and later through shell burning above an electron-degenerate carbon-oxygen core. For massive (more than ten solar masses, $>10 M_{\odot}$) stars, direct nuclear burning continues with the production of oxygen, neon, magnesium, silicon and so on, cumulating in the synthesis of iron, the heaviest element possible through direct nuclear burning. The other heavy elements, from yttrium and zirconium to uranium and beyond, are produced by neutron capture followed by β decay¹.

For the majority of stars ($\sim 95\%$, corresponding to stars with initial masses of less than $\sim 8 M_{\odot}$), direct nuclear burning does not proceed beyond helium, and carbon is never ignited. Most of the nucleosynthesis occurs through slow neutron capture (the s process) during the asymptotic giant branch (AGB), a brief phase ($\sim 10^6$ yr) of stellar evolution where hydrogen and helium burn alternately in a shell. These newly synthesized elements are raised to the surface through periodic 'dredge-up' episodes, and the observation of short-lived isotopes in stellar atmospheres provides direct evidence that nucleosynthesis is occurring in AGB stars².

However, the atomic form is not the only form in which chemical elements can occur. Molecules are major constituents of planetary atmospheres and the Solar System contains many solid bodies, such as asteroids and terrestrial planets. Since the late 1960s, advances in millimetre-wave observation have led to the detection of over 120 molecules in the interstellar medium. Although a variety of chemical models have been developed to explain how molecular synthesis could occur in the interstellar medium, we have no direct observational knowledge of the timescales on which these reactions occur. The discovery of a large number of gas-phase molecules as well as submicrometre-sized solid-state particles of various chemical compositions in the circumstellar envelopes of AGB stars has led to the realization that extensive molecular synthesis occurs during this brief phase of stellar evolution. These envelopes are expanding and have dynamical timescales of several thousand years only, providing the first definite indication that production of molecules and solids of high complexity can occur rapidly in the circumstellar environment.

Although it was commonly assumed that stellar grains are destroyed during their journey through the interstellar medium by radiation and shocks³, the recent discoveries of the chemical richness of circumstellar grains have raised the possibility that

circumstellar molecular synthesis may have significant implications for the chemical enrichment of the Galaxy, or even of the early Solar System. The exploration of this new perspective is the subject of this review.

The asymptotic giant branch and beyond

The approximately ten thousand years of evolution following the end of the AGB phase represents a most fascinating 'laboratory' for astrochemistry. A large variety of organic species have been observed to emerge in the circumstellar environment of stars evolving through this stage, following the formation of gas-phase molecules and inorganic solid-state compounds in the previous AGB phase. The well-defined conditions of the environment approximate a controlled laboratory as closely as is possible in space, providing severe constraints on any model of gas-phase and solid-state chemistry.

Stars on the galactic disk (population I) begin their AGB phase with more oxygen than carbon. As more and more carbon is made in the core and dredged up to the surface, the abundance of carbon in the stellar photosphere will eventually exceed that of oxygen. Most of the carbon will combine with oxygen to form the stable molecule carbon monoxide (CO). The excess carbon atoms will be available to form other carbon-based molecules such as C_2 , C_3 and CN. The presence of these molecular bands in the photospheric spectrum distinguishes these stars from normal oxygen-rich AGB stars, and stars that show these spectral characteristics are called carbon stars.

As an AGB star evolves, both its luminosity and size increase. A typical AGB star has a luminosity several thousand times that of the Sun, and a radius several hundred times the solar radius. The combination of high luminosity and large size causes envelope pulsation, and when aided by radiation pressure, an AGB star will eject its outer layers in the form of a stellar wind, with ejection rates as high as $10^{-4} M_{\odot} \text{ yr}^{-1}$ (ref. 4). A star can therefore lose a major fraction of its original mass through these stellar winds, creating a thick, dense envelope around the central star.

Because of the low temperatures, most of the gas in the circumstellar envelopes is in molecular form, and can be detected through their rotational or vibrational transitions. In the past 30 years, millimetre-wave and near-infrared spectroscopy has detected the rotational and vibrational stretching mode transitions of ~ 60 molecules in the circumstellar envelopes of AGB stars. The detected molecular species include inorganics (CO, SiO, SiS, NH_3 , AlCl), organics (C_2H_2 , CH_4 , H_2CO , CH_3CN), radicals (CN, C_2H , C_3 , HCO^+), cyclic molecules (C_3H_2), and cyanopolynes (HCN, HC_3N , ... HC_9N) (ref. 5).

In addition to molecules, solid-state species, both amorphous and crystalline, are also found in the circumstellar envelopes of AGB stars. The chemical composition of solids can be identified through their lattice vibrational modes. The most common solid-state condensate is amorphous silicates and silicon carbide (SiC). These solid particles are believed to condense directly from gas-phase molecules, as the gas temperature cools during the envelope expansion⁶.

Because solid particles have a high opacity to visible light, the condensation in the circumstellar envelope can cause large obscuration to the light from the central star. The intercepted starlight heats up the dust particles, which in turn cools by self-radiating in the infrared. As the ejection rate increases in the late stages of AGB evolution, the envelope can contain so much dust that the star is completely obscured by dust extinction. Such extremely evolved AGB stars have no optical counterpart and can only be identified by infrared observations (Fig. 1).

The high rate of mass loss on the AGB will eventually deplete the envelope of the star and gradually expose the hot core. As the envelope thins to below $10^{-3} M_{\odot}$, the star will begin to increase in temperature. Through a combination of hydrogen shell burning and mass loss, the envelope becomes thinner, leading to an increase of stellar temperature. When the stellar temperature reaches 25,000 K, the ultraviolet photon output from the star will begin to photo-ionize the circumstellar envelope, creating a planetary nebula (PN). The intervening phase between the end of AGB and the onset of photo-ionization is called the proto-planetary nebula (PPN) phase^{7,8}.

Although PNe are bright visible objects because of their emission-line spectrum, recent observations of PNe in the infrared and millimetre wavelengths have shown that they also possess thick molecular envelopes mixed with solid-state particles. Mid- and far-infrared observations from the Infrared Astronomical Satellite (IRAS) and Infrared Space Observatory (ISO) missions have found that dust emission represents a major fraction of energy output from PNe⁹. The detection of molecular/dust envelopes in PNe clearly establishes the link to their progenitor AGB stars.

By observing the spectral changes of the envelope as a function of evolution between AGB stars and PNe, we are able directly to

determine the chemical steps through which complex organic and inorganic compounds are made in the circumstellar environment. In comparison to interstellar clouds, there are many advantages to using circumstellar envelopes to study the process of chemical synthesis. First, there is only one single energy source—the central star, whose temperature and luminosity are well known. The envelope often has a well-defined symmetry, which simplifies the geometry of the system. By using molecular lines and infrared continuum radiation as probes, the physical conditions of the envelope, such as density, temperature and radiation background, are well determined as a function of distance from the star. Most importantly, the chemical reaction times are constrained by the dynamical and stellar evolution times, which are $\sim 10^4$ yr for AGB stars, 10^3 yr for PPNe, and 10^4 yr for PNe. These timescales therefore give us a precise knowledge of the chemical time needed to form one species from another and thus rigorously constrain the chemical models^{10,11}.

Infrared features of inorganic compounds

By comparing the spectral features of known minerals with astronomical spectroscopic observations, we are able to identify inorganic solid-state materials that are synthesized in stars. However, the circumstellar envelopes of stars may harbour new substances made under conditions different from those of the terrestrial environment. Through the discovery of new spectral features and the subsequent identification of their carriers, we may learn about new chemical processes or materials that have not previously been considered in the laboratory setting.

The kinds of molecules and solid particles that form in the circumstellar envelopes of AGB stars are dependent on the chemical composition of the stars. Because CO molecules take up most of the elemental form of whichever is less abundant of oxygen and carbon, oxides are generally found in oxygen-rich AGB stars, whereas organic compounds are generally limited to carbon-rich AGB stars.

The first solid-state condensate detected in the circumstellar envelopes of AGB stars is composed of silicates. Because silicates make up the most common group of minerals in the Earth's crust and both oxygen and silicon are among the most abundant elements in the Universe, it is reasonable to expect that silicates of stellar origin may also be common. Although most of the terrestrial silicates are in crystalline form, most of the silicates found in AGB stars are amorphous. This is consistent with the expectation that amorphous structures result from rapid condensation from the gas phase in AGB winds. The 9.7- and 18- μ m features of amorphous silicates are detected in over 4,000 oxygen-rich AGB stars¹². During the ISO mission, a rich family of narrow emission bands was found in the spectra of AGB stars and PNe and they are identified as originating from crystalline silicates¹³. Crystalline silicates can be in the form of olivine ($Mg_{2-2x}Fe_{2x}SiO_4$, where $0 \leq x \leq 1$) or pyroxene ($Mg_{1-x}Fe_xSiO_3$). Analyses of the ISO observations suggest that the crystalline silicates in AGB stars are magnesium-rich ($x \approx 0$). From the strengths of the crystalline silicate features we can estimate that the abundance of crystalline silicates is small in comparison to amorphous silicates.

In addition to silicates, a variety of refractory oxides may also be present in circumstellar envelopes of AGB stars. A broad feature centred around 13 μ m, first discovered by the IRAS Low Resolution Spectrometer (LRS)¹⁴, could be due to corundum (α - Al_2O_3 , which has a feature at 12.7 μ m), glass (amorphous SiO_2 , 12.3 μ m), spinel (12.95 μ m), or rutile (13.4 μ m)¹⁵. An emission feature at 19.5 μ m observed by ISO has been attributed to mixture of Mg-Fe oxides (such as $Mg_{0.1}Fe_{0.9}O$)¹⁶.

In carbon-rich AGB stars, the most common solid-state condensate is SiC, whose 11.3- μ m feature due to the stretch mode of the Si-C bond is detected in emission in 700 carbon stars¹². In more evolved carbon stars, the SiC feature becomes weaker, suggesting that the dust constituent is increasingly dominated by amorphous carbon.

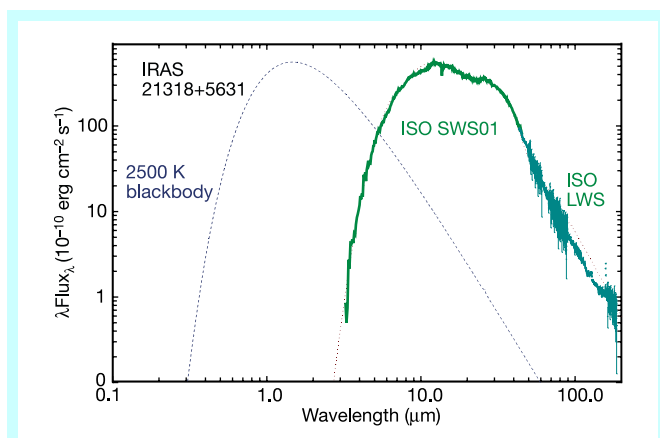


Figure 1 Invisible stars. IRAS 21318+5631 is an example of an evolved carbon star so obscured by its own ejected circumstellar dust envelope that the central star suffers from 360 magnitudes of extinction in the visible (A_V), and is undetectable in the optical region. Its infrared spectrum (solid lines: green, ISO SWS01; turquoise, ISO LWS) is completely due to dust emission and has a colour temperature of 300 K. The red dotted line represents the theoretical fit to the spectrum based on a one-dimensional radiation transfer model with a hidden 2,500 K central star (blue dashed line) as the energy source. The absorption feature near the peak of the spectrum is the 13.7- μ m band of acetylene. Infrared and millimetre-wave spectroscopic observations have shown that these stars are prolific molecular factories.

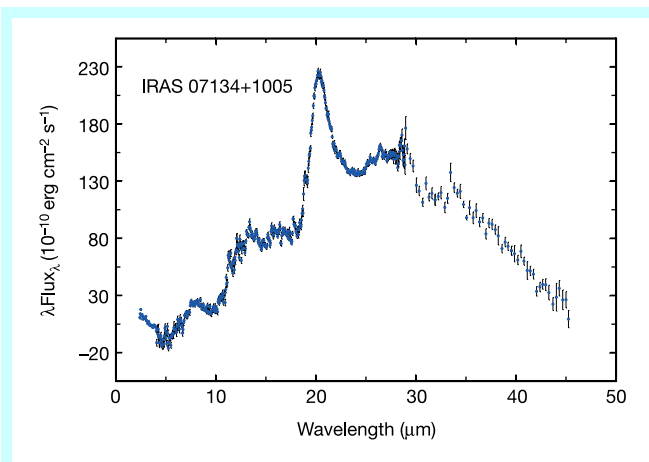


Figure 2 The unidentified emission feature at 21 μm . ISO SWS spectrum of the PPN IRAS 07134+1005 showing the very strong 21- μm emission feature. The only reason this feature was not detected earlier is because it has been difficult to do mid-infrared spectroscopy from ground-based telescopes and because it is only present in PPNs, a new class of celestial object with very short lifetimes. The error bars are 1σ deviation of data points from different scans and detectors over each wavelength resolution element.

In addition to spectral features that can be identified as originating from known minerals, there are strong emission features that remain unidentified. The 21- μm feature was first discovered in four PPNs from observations by the IRAS LRS¹⁷ (Fig. 2). As of 2004, there are 12 known 21- μm sources; all are carbon-rich stars in the post-AGB phase of evolution¹⁸. The fact that all of the 21- μm sources are carbon-rich strongly suggests that the carrier of this feature is carbon-based. High-resolution ($\Delta\lambda/\lambda \approx 2,000$) ISO observations have found that all features have the same intrinsic profile and peak wavelength ($\lambda = 20.1 \mu\text{m}$)¹⁹. There is no evidence for any discrete substructure caused by molecular bands in the

observed spectra, suggesting that the 21- μm feature is either due to a solid substance or a mixture of many similarly structured large molecules. Possible candidates that have been proposed include large polycyclic aromatic hydrocarbon (PAH) clusters, hydrogenated amorphous carbon (HAC) grains²⁰, hydrogenated fullerenes²¹, nanodiamonds²² and TiC nanoclusters²³, but none of these identifications has been found to be satisfactory. The recent suggestion that the feature is due to micrometre-sized β SiC or nano-SiC grains with carbon impurities may be the most plausible²⁴, although this hypothesis requires an efficient suppression of the 11.3- μm feature.

Another unidentified solid-state circumstellar feature is a strong feature at around 30 μm . This feature was first discovered in the spectra of carbon stars and PNe from the Kuiper Airborne Observatory (KAO) observations²⁵. More recently, the 30- μm feature has been found to be common in carbon-rich PPNs, especially those showing the 21- μm emission feature²⁶. The fact that a significant fraction ($\sim 20\%$) of the total radiative energy output of PPNs is emitted in this feature suggests that the carrier must be composed of abundant elements²⁷. The first suggested identification was solid MgS based on a comparison with laboratory measurements²⁸. The alternative suggestion that the carrier is a carbonaceous material continues to be popular because the feature is seen only in carbon-rich objects.

The synthesis of organic compounds

Although many of the gas-phase molecules detected through their rotational transitions are organic molecules, the detection of complex organic compounds through their stretching and bending vibrational modes was totally unexpected. These organic compounds have dozens or hundreds of atoms, and their synthesis was not thought to be possible in a low-density environment. The detection of organic compounds in stellar winds gives us important information on how these species are formed. The evolution from AGB star to PPN to PN is very short, and this gives us precise knowledge of the timescale of chemical synthesis.

A family of strong infrared emission features at 3.3, 6.2, 7.7, 8.6, 11.3 and 12.7 μm were first detected by the KAO in the young carbon-rich PN NGC 7027²⁹, and have since been widely observed

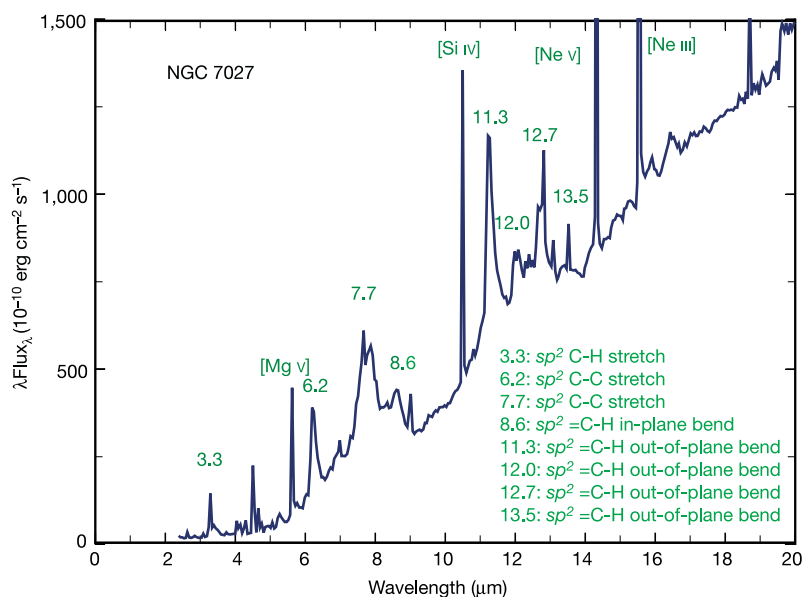


Figure 3 Stretching and bending modes of aromatic compounds. The ISO SWS spectrum of the PN NGC 7027 is dominated by a strong dust continuum (peaking at $\sim 30 \mu\text{m}$), and atomic emission lines due to fine-structure transitions of heavy

elements (for example, the [Ne III] line at 15.6 μm). The AIB features are marked by their peak wavelengths and their identification is listed in the legend.

in PNe, H II regions, reflection nebulae and galaxies (Fig. 3). These features are identified with the aromatic C–H stretch ($3.3\ \mu\text{m}$), aromatic C–C stretch (6.2 and $7.7\ \mu\text{m}$), aromatic C–H in-plane bend ($8.6\ \mu\text{m}$), and aromatic C–H out-of-plane bending ($11.3\ \mu\text{m}$) modes³⁰. These features are collectively referred to as aromatic infrared bands (AIB). Since the AIB features first emerge in the PPN phase, what are the steps leading to the formation of ring molecules? Acetylene (C_2H_2), believed to be the first building block of benzene, is commonly detected in evolved carbon stars through its ν_5 fundamental band at $13.7\ \mu\text{m}$ (see Fig. 1). Polymerization of C_2H_2 leads to the formation of diacetylene (C_4H_2) and triacetylene (C_6H_2) in PPN, culminating in the formation of benzene³¹.

In addition to the AIBs seen in PNe, emission features at 3.4 and $6.9\ \mu\text{m}$ due to aliphatic C–H stretch and bending modes are found in the spectra of PPNe and young PNe^{20,27,32,33}. Furthermore, emission features at 11.3 , 12.1 , 12.4 and $13.3\ \mu\text{m}$, which can be identified as arising from out-of-plane vibrational modes of aromatic C–H bonds with respectively one, two, three or four C–H bonds per edge of an aromatic ring³⁴, have also been detected³⁵.

Aromatic and aliphatic bands are not the only emission features seen in the infrared spectrum of PPNe. Also present are broad emission features at 8 and $12\ \mu\text{m}$ (Fig. 4). Because the $6.9\text{-}\mu\text{m}$ band is known to originate from a mixture of $-\text{CH}_2-$ and $-\text{CH}_3$ bending modes, associated bending modes of other side groups are likely to be present. Similarly, the $11.3\text{-}\mu\text{m}$ aromatic C–H out-of-plane bending mode can be accompanied by a complex set of features due to out-of-plane vibrations of alkenes³⁶. The existence of the 8 - and $12\text{-}\mu\text{m}$ broad emission features therefore suggests that the chemical structures of these carbonaceous compounds are complex, and probably include a variety of alkane and alkene side groups attached to aromatic rings.

The discovery of aromatic features has led to vigorous discussions in the literature on the nature of the carriers. Most of the suggestions centre on aromatic hydrocarbons of both natural and artificially created varieties. Carbonaceous compounds have been produced in the laboratory by a variety of techniques, including chemical vapour deposition³⁷, combustion³⁸, laser ablation³⁹, and laser pyrolysis of

hydrocarbons⁴⁰. The most widely discussed candidates for the carrier of the AIB are polycyclic aromatic hydrocarbon (PAH) molecules. PAH molecules are benzene rings of sp^2 -hybridized C atoms linked to each other in a plane, with H atoms or other radicals saturating the outer bonds of peripheral C atoms. PAH molecules were first proposed as the carrier of the AIBs in the diffuse interstellar medium because their small sizes ($<1\ \text{nm}$) allow them to be heated to high temperatures ($\sim 1,000\ \text{K}$) by stochastic heating^{41–43}.

However, small gas-phase PAH molecules are unlikely to be the carrier of the AIBs seen in PPNe and PNe. The AIBs in these objects always sit on top of a strong continuum, which cannot be provided by small PAH molecules. PPNe have no ultraviolet photon output, which would be required for the excitation of gas-phase PAHs. The width of the AIB features and the consistency of the peak wavelengths also argue against the idea that they are produced by a mixture of different PAH molecules. The carrier is more likely to be a solid-state compound, consisting of at least hundreds of carbon atoms.

The detection of aliphatic signatures in PPNe has greatly increased the interest in another natural substance: coal. Formed from fossilized hydrocarbon materials, coal contains a mixture of sp , sp^2 and sp^3 bonds. The possibility that coal-like material (for example, kerogen) can be responsible for the AIB was first suggested by R. Papoular⁴⁴. Structurally, kerogen can be represented by random arrays of aromatic carbon sites, aliphatic chains ($-\text{CH}_2-$)_n, and linear chains of benzenic rings with functional groups made up of H, O, N and S attached. Stellar synthesis of organic compounds is unlikely to be pure, involving only H and C, so the presence of other elements in carbonaceous material is to be expected. Comparison of the infrared spectra of kerogen shows that they are very similar to those of PPNe^{45–47}.

The AIBs can also arise from artificial substances not common in the natural terrestrial environment. Examples of such compounds include hydrogenated amorphous carbon (HAC) and quenched carbonaceous composites (QCC). HAC consists of islands of aromatic (sp^2) bonded C atoms joined together with a variety of

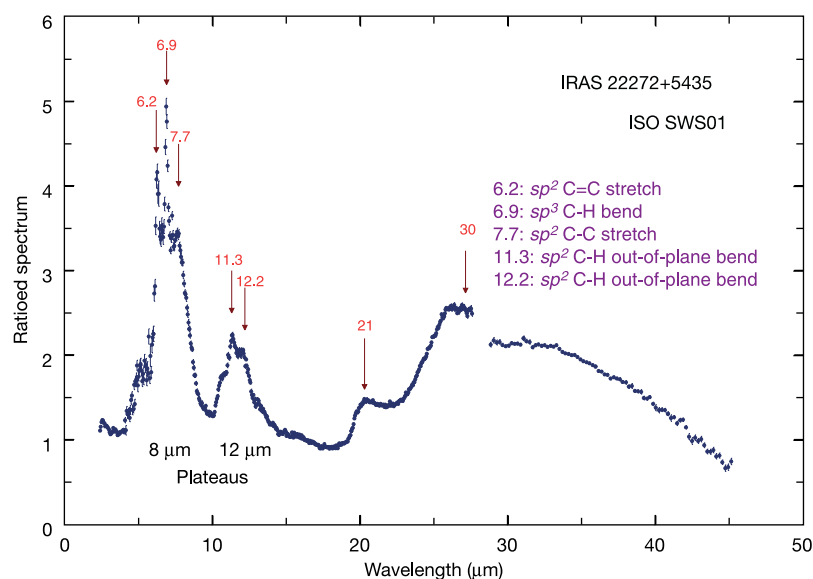


Figure 4 Broad emission features in PPNe. This continuum-removed ISO SWS01 spectrum of the PPN IRAS 22272+5435 shows the 8 - and $12\text{-}\mu\text{m}$ emission plateaus due to the in-plane and off-plane bending modes of aliphatic side groups attached to an aromatic carbonaceous compound, for which the narrow emission features and

their peak wavelengths are marked on the spectrum. Also shown are the unidentified emission features at 21 and $30\ \mu\text{m}$ (ref. 36). The error bars are 1σ deviation of data points from different scans and detectors over each wavelength resolution element.

peripheral sp^2 and sp^3 bonded hydrocarbons⁴⁸. HAC can be formed in the laboratory by direct condensation of carbon vapour from an H-rich atmosphere, and has been observed to show similar spectral features as in PPNe⁴⁹. QCC are produced by the technique of hydrocarbon plasma deposition. Methane gas is heated to 3,000 K with a microwave generator, allowed to expand into a vacuum chamber and condensed on a room-temperature substrate⁵⁰. The resultant dark, granular material is shown by electron micrography to have an amorphous structure. Mass spectroscopy of QCC suggests that their aromatic component typically consists of one to four rings, and most have only one to two rings. Infrared spectroscopy of QCC reveals a mixture of sp , sp^2 and sp^3 bonds. The conditions under which QCC are synthesized resemble the circumstellar environment of AGB stars, so its laboratory production may shed some light on the solid-state condensation process in the envelopes of AGB and post-AGB stars.

Although it is not yet certain whether coal, HAC or QCC gives the best approximation to interstellar carbonaceous dust, it is clear that such grains include both aromatic and aliphatic components. A schematic of possible chemical structure is given in Fig. 5.

Stardust in the Solar System

Spectral signatures of grains produced in AGB stars, PPNe and PNe are now found in Solar System objects. These results suggest that comets, interplanetary dust and meteorites preserve pristine stellar material not processed by the early solar nebula, and our ability to

perform laboratory analysis of these materials through mass spectroscopic and isotopic analysis has given us a direct link to inorganic and organic compounds produced by stars.

The silicate features, both amorphous and crystalline, are seen in comets⁵¹, interplanetary dust⁵² and meteorites^{53,54}. Infrared spectra of organic extract sublimate from the Murchison meteorite show the 3.4- μm aliphatic features that are similar to those observed in PPNe⁵⁵. This feature has also been observed in comets⁵⁶ and interplanetary dust⁵⁷. Isotopic studies of meteorites have also identified grains of presolar origin, including diamonds⁵⁸, corundum⁵⁹, spinel⁵⁹, and in particular SiC⁶⁰, the most common and best-studied of the group. These grains therefore represent an important link between stars and the Solar System⁶¹. The dominant organic matter in carbonaceous chondrites is similar to kerogen⁶² and aliphatic sidegroups constitute several per cent of the mass of interplanetary dust⁶³.

Outlook

We have learned that chemical synthesis of complex organic and inorganic compounds can take place in the low-density circumstellar environment over very short (10^3 yr) timescales during the late stages of stellar evolution. By tracing the change in the infrared spectra of carbon stars to PPNe to PNe, we found definite evidence for chemical evolution (Table 1). Infrared spectroscopic observations have clearly demonstrated that solid-state compounds of both organic and inorganic nature are produced in abundance in

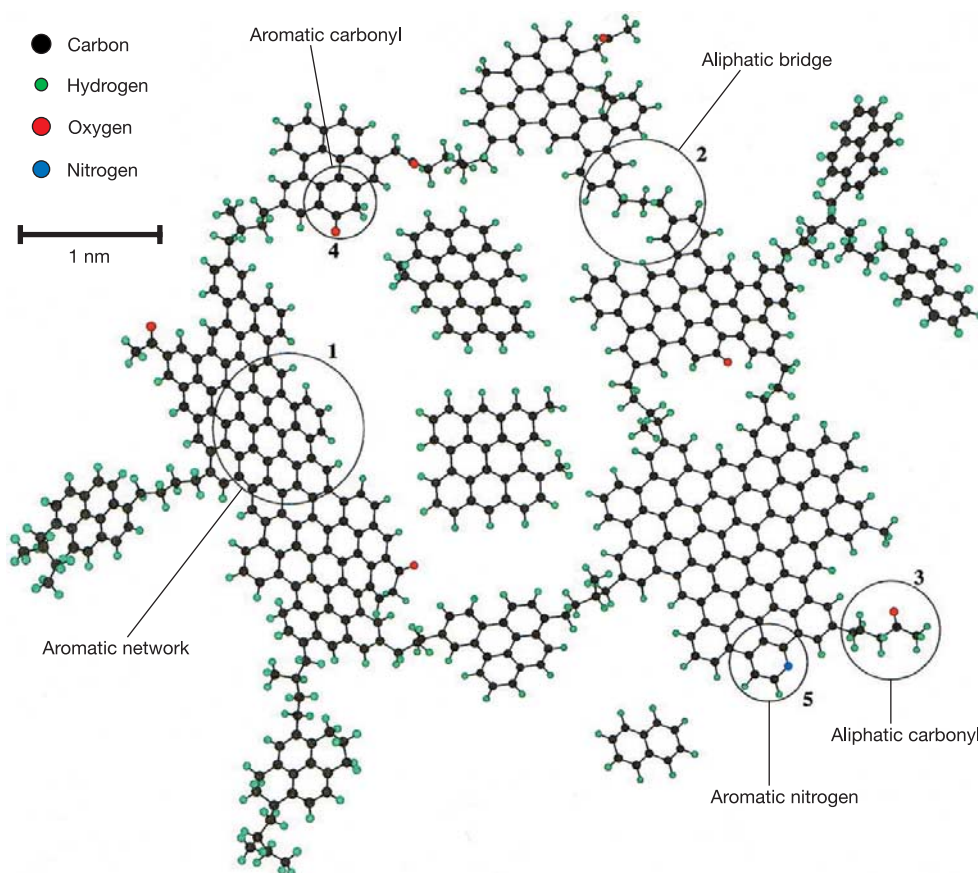


Figure 5 A model of carbonaceous interstellar dust consisting of various aromatic and aliphatic groups (after ref. 55). Examples of five basic units are shown, circled and labelled. The carbonaceous compounds seen in PPNe and PNe are likely to include both aromatic and aliphatic components as well as impurities such as O and N. Among the 470 C atoms in this structure, 400 are in aromatic units and 70 are in aliphatic

groups. The approximate volume of this structure is $\sim 10^{-19} \text{ cm}^3$, that is, approximately 10^4 of such structures could be contained in a typical grain of 0.1 μm . The exact ratios of sp^2 and sp^3 sites will vary, depending on the state of evolution of the central star.

Table 1 Changes in relative strengths of infrared emission features as stars evolve from AGB to planetary nebulae

Infrared features (μm)	Origin	Carbon stars	Proto-planetary nebulae	Planetary nebulae
3.3, 6.2, 7.7, 11.3	Aromatic stretch and bending modes	No	Yes	Strong
3.4, 6.9	C–H aliphatic stretch and bend	No	Yes	Weak
12.1, 12.4, 13.3	C–H out-of-plane bend with two, three and four adjacent H atoms	No	Yes	Weak
Broad 8, 12	Bending modes from aliphatic sidegroups	No	Yes	Weak
Broad 21	–	Weak	Strong	No
Broad 30	–	Yes	Yes	Yes

The weakening of the 3.4 and 6.9 μm features from PPNe to PNe suggests a change from aliphatic to aromatic structures. This could be the result of photochemistry where the onset of ultraviolet radiation modifies the aliphatic side groups through isomerizations, bond migrations, cyclization and ring closures and transforms them into ring systems³⁶. Hydrogen loss can also result in fully aromatic rings that are more stable than alkanes or alkenes. Evidence for such H loss can also be found in the weakening of the 12.1, 12.4 and 13.3 μm features and the strengthening of the 11.3 μm feature from PPNe to PNe.

the circumstellar envelopes of evolved stars. The ejection of these grains into the interstellar medium suggests that circumstellar grains are a major source of interstellar grains, some of which have even survived the formation of the Solar System. The detections of pre-solar grains in meteorites and the detection of a high deuterium to hydrogen ratio in interplanetary dust particles⁶⁴ have supported this stellar–Solar System connection. Space missions to comets allow *in situ* analysis of dust compositions in comets through mass spectroscopy, yielding both size and elemental abundance determinations. Instead of being condensates out of a homogeneous solar nebula that has vaporized all pre-solar material, we now believe that the Solar System objects contain primordial materials that have been delivered to the solar nebula from stars. The possibility of sample gathering from interplanetary space and comets will allow us to obtain specific information on the structure of the grains and sheds light on the process of nucleation and condensation from the gas phase under non-equilibrium conditions.

With the analysis of meteoritic material and sample returned from comets and in the interplanetary medium⁶⁵, we are now able to hold materials condensed in the stellar environment in our own hands, and perform a variety of analyses using all the laboratory techniques at our disposal. With techniques such as scanning transmission X-ray microscopy⁶⁶, high-resolution transmission electron microscopy, nuclear magnetic resonance spectroscopy⁶⁷, and ion microprobe scanning imaging⁶⁸, we can make high spatial resolution measurements of the physical and chemical properties of stellar grains. These techniques allow precise determination of the geometry, size and shape of the aliphatic groups, as well as their exact chemical nature. All these properties are difficult to obtain from spectroscopy alone. Comparison of the actual isotopic ratios in these grains to their predicted values by theoretical nucleosynthesis models will help to constrain the nuclear processes in the interior of stars.

In the coming decade, laboratory astrophysics will expand beyond its traditional role of comparing laboratory measurements and astronomical observations to include the analysis of stellar-synthesized materials. For most of the history of the science, astronomy has been restricted to remote observations. Through a comparison between infrared spectroscopy of stellar ejecta and laboratory analysis of collected pre-solar grains, we have entered a new era of direct testing of theories of nucleosynthesis in the interior of stars and chemical synthesis of inorganic and organic compounds in the circumstellar environment. □

doi:10.1038/nature02862.

1. Wallerstein, G. et al. Synthesis of the elements in stars: forty years of progress. *Rev. Mod. Phys.* **69**, 995–1084 (1997).
2. Lattanzio, J. in *Planetary Nebulae: Their Evolution and Role in the Universe* (eds Kwok, S., Dopita, M. & Sutherland, R.) 73–81 (ASP, San Francisco, 2003).
3. Jones, A. P., Tielens, A. G. G. M., Hollenbach, D. J. & McKee, C. F. Grain destruction in shocks in the interstellar medium. *Astrophys. J.* **433**, 797–810 (1994).
4. Kwok, S. Effects of mass loss on the late stages of stellar evolution. *Phys. Rep.* **156**, 113–146 (1987).
5. Olofsson, H. The neutral envelopes around AGB and post-AGB objects. *IAU Symp. 178: Molecules in Astrophysics: Probes and Processes* (ed. van Dishoeck, E.) 457–468 (Kluwer, Dordrecht, 1997).

6. Lodders, K. & Fegley, B. The origin of circumstellar silicon carbide grain found in meteorites. *Meteoritics* **30**, 661–678 (1995).
7. Kwok, S. Proto-planetary nebulae. *Annu. Rev. Astron. Astrophys.* **31**, 63–92 (1993).
8. Van Winckel, H. Post-AGB Stars. *Annu. Rev. Astron. Astrophys.* **41**, 391–427 (2003).
9. Zhang, C. Y. & Kwok, S. Spectral energy distributions of compact planetary nebulae. *Astron. Astrophys.* **250**, 179–211 (1991).
10. Woods, P. M., Millar, T. J., Herbst, E. & Zijlstra, A. A. The chemistry of protoplanetary nebulae. *Astron. Astrophys.* **402**, 189–199 (2003).
11. Hasegawa, T., Volk, K. & Kwok, S. A chemical model of the neutral envelope of the planetary nebula NGC 7027. *Astrophys. J.* **532**, 994–1005 (2000).
12. Kwok, S., Volk, K. & Bidelman, W. P. Classification and identification of IRAS sources with low-resolution spectra. *Astrophys. J. Suppl.* **112**, 557–584 (1997).
13. Jäger, C. et al. Steps toward interstellar silicate mineralogy. IV. The crystalline revolution. *Astron. Astrophys.* **339**, 904–916 (1998).
14. Sloan, G. C., LeVan, P. D. & Little-Marein, I. R. Sources of the 13 μm feature associated with oxygen-rich circumstellar dust. *Astrophys. J.* **463**, 310–319 (1996).
15. Posch, T. et al. On the origin of the 13 μm feature. A study of ISO-SWS spectra of oxygen-rich AGB stars. *Astron. Astrophys.* **352**, 609–618 (1999).
16. Posch, T., Kerschbaum, F., Mutschke, H., Dorschner, J. & Jäger, C. On the origin of the 19.5 μm feature. Identifying circumstellar Mg-Fe-oxides. *Astron. Astrophys.* **393**, L7–L10 (2002).
17. Kwok, S., Volk, K. & Hrivnak, B. J. A 21 micron emission feature in four proto-planetary nebulae. *Astrophys. J.* **345**, L51–L54 (1989).
18. Kwok, S., Volk, K. & Hrivnak, B. J. in *IAU Symp. 191: Asymptotic Giant Branch Stars* (eds Le Bertre, T., Lèbre, A. & Waelkens, C.) 297–302 (ASP, San Francisco, 1999).
19. Volk, K., Kwok, S. & Hrivnak, B. J. High-resolution Infrared Space Observatory spectroscopy of the unidentified 21 micron feature. *Astrophys. J.* **516**, L99–L102 (1999).
20. Buss, R. H. Jr et al. Hydrocarbon emission features in the IR spectra of warm supergiants. *Astrophys. J.* **365**, L23–L26 (1990).
21. Webster, A. The lowest of the strongly infrared active vibrations of the fullerenes and astronomical emission band at a wavelength of 21-microns. *Mon. Not. R. Astron. Soc.* **277**, 1555–1566 (1995).
22. Hill, H. G. M., Jones, A. P. & d'Hendecourt, L. B. Diamonds in carbon-rich proto-planetary nebulae. *Astron. Astrophys.* **336**, L41–L44 (1998).
23. von Helden, G. et al. Titanium carbide nanocrystals in circumstellar environments. *Science* **288**, 313–316 (2000).
24. Speck, A. K. & Hofmeister, A. M. Processing of presolar grains around post-asymptotic giant branch stars: silicon carbide as the carrier of the 21 micron feature. *Astrophys. J.* **600**, 986–991 (2004).
25. Forrest, W. J., Houck, J. R. & McCarthy, J. F. A far-infrared emission feature in carbon-rich stars and planetary nebulae. *Astrophys. J.* **248**, 195–200 (1981).
26. Omont, A. et al. The 30 micron emission band in carbon-rich pre-planetary nebulae. *Astrophys. J.* **454**, 819–825 (1995).
27. Hrivnak, B. J., Volk, K. & Kwok, S. 2–45 micron infrared spectroscopy of carbon-rich proto-planetary nebulae. *Astrophys. J.* **535**, 275–292 (2000).
28. Goebel, J. H. & Moseley, S. H. Laboratory infrared spectra of predicted condensates in carbon-rich stars. *Astrophys. J.* **290**, L35–L39 (1985).
29. Russell, R. W., Soifer, B. T. & Willner, S. P. The 4 to 8 micron spectrum of NGC 7027. *Astrophys. J.* **217**, L149–L153 (1977).
30. Duley, W. W. & Williams, D. A. The infrared spectrum of interstellar dust - Surface functional groups on carbon. *Mon. Not. R. Astron. Soc.* **196**, 269–274 (1981).
31. Cernicharo, J. et al. Infrared Space Observatory's discovery of C_4H_2 , C_6H_2 , and benzene in CRL 1018. *Astrophys. J.* **546**, L123–L126 (2001).
32. Jourdain de Muizon, M., d'Hendecourt, L. B. & Geballe, T. R. Three micron spectroscopy of IRAS sources—Observed and laboratory signatures of PAHs. *Astron. Astrophys.* **235**, 367–378 (1990).
33. Geballe, T. R., Tielens, A. G. G. M., Kwok, S. & Hrivnak, B. J. Unusual 3 micron emission features in three proto-planetary nebulae. *Astrophys. J.* **387**, L89–L91 (1992).
34. Hony, S. et al. The CH out-of-plane bending modes of PAH molecules in astrophysical environments. *Astron. Astrophys.* **370**, 1030–1043 (2001).
35. Kwok, S., Volk, K. & Hrivnak, B. J. Chemical evolution of carbonaceous materials in the last stages of stellar evolution. *Astron. Astrophys.* **350**, L35–L38 (1999).
36. Kwok, S., Volk, K. & Bernath, P. On the origin of infrared plateau features in proto-planetary nebulae. *Astrophys. J.* **554**, L87–L90 (2001).
37. Sakata, A., Wada, S., Onaka, T. & Tokunaga, A. T. Infrared spectrum of quenched carbonaceous composite (QCC). II—A new identification of the 7.7 and 8.6 micron unidentified infrared emission bands. *Astrophys. J.* **320**, L63–L67 (1987).
38. Colangeli, L., Mennella, V., Palumbo, P., Rotundi, A. & Bussolletti, E. Mass extinction coefficients of various submicron amorphous carbon grains: Tabulated values from 40 nm to 2 μm . *Astron. Astrophys. Suppl.* **113**, 561–577 (1995).
39. Scott, A. & Duely, W. W. The decomposition of hydrogenated amorphous carbon: a connection with polycyclic aromatic hydrocarbon molecules. *Astrophys. J.* **472**, L123–L125 (1996).

40. Herlin, N. *et al.* Nanoparticles produced by laser pyrolysis of hydrocarbons: analogy with carbon cosmic dust. *Astron. Astrophys.* **330**, 1127–1135 (1998).
41. Puget, J. L. & Léger, A. A new component of the interstellar matter: small grains and large aromatic molecules. *Annu. Rev. Astron. Astrophys.* **27**, 161–198 (1989).
42. Allamandola, L. J., Tielens, A. G. G. M. & Barker, J. R. Interstellar polycyclic aromatic hydrocarbons—the infrared emission bands, the excitation/emission mechanism, and the astrophysical implications. *Astrophys. J. Suppl.* **71**, 733–775 (1989).
43. Sellgren, K. Aromatic hydrocarbons, diamonds, and fullerenes in interstellar space: puzzles to be solved by laboratory and theoretical astrochemistry. *Spectrochim. Acta A* **57**, 627–642 (2001).
44. Papoular, R., Conard, J., Giuliano, M., Kister, J. & Mille, G. A coal model for the carriers of the unidentified IR bands. *Astron. Astrophys.* **217**, 204–208 (1989).
45. Guillois, O., Nenner, I., Papoular, R. & Reynaud, C. Coal models for the infrared emission spectra of proto-planetary nebulae. *Astrophys. J.* **464**, 810–817 (1996).
46. Papoular, R. *et al.* A comparison of solid-state carbonaceous models of cosmic dust. *Astron. Astrophys.* **315**, 222–236 (1996).
47. Papoular, R. The use of kerogen data in understanding the properties and evolution of interstellar carbonaceous dust. *Astron. Astrophys.* **378**, 597–607 (2001).
48. Jones, A. P., Duley, W. W. & Williams, D. A. The structure and evolution of hydrogenated amorphous carbon grains and mantles in the interstellar medium. *Q. J. R. Astron. Soc.* **31**, 567–582 (1990).
49. Scott, A. D., Duley, W. W. & Jahani, H. R. Infrared emission spectra from hydrogenated amorphous carbon. *Astrophys. J.* **490**, L175–L177 (1997).
50. Sakata, A., Wada, A., Tanab'e, T. & Onaka, T. Infrared spectrum of the laboratory-synthesized quenched carbonaceous composite (QCC): comparison with the infrared unidentified emission bands. *Astrophys. J.* **287**, L51–L54 (1984).
51. Wooden, D. H. *et al.* Silicate mineralogy of the dust in the inner coma of comet C/1995 01 (Hale-Bopp) pre- and postperihelion. *Astrophys. J.* **517**, 1034–1058 (1999).
52. Messenger, S., Keller, L. P., Stadermann, F. J., Walker, R. M. & Zinner, E. Samples of stars beyond the solar system: silicate grains in interplanetary dust. *Science* **300**, 105–108 (2003).
53. Nagashima, K., Krot, A. N. & Yurimoto, H. Stardust silicates from primitive meteorites. *Nature* **428**, 921–924 (2004).
54. Nguyen, A. N. & Zinner, E. Discovery of ancient silicate stardust in a meteorite. *Science* **303**, 1496–1499 (2004).
55. Pendleton, Y. J. & Allamandola, L. J. The organic refractory material in the diffuse interstellar medium: mid-infrared spectroscopic constraints. *Astrophys. J. Suppl.* **138**, 75–98 (2002).
56. Davies, J. K. *et al.* The infrared (3.2–3.6 micron) spectrum of Comet P/Swift-Tuttle—Detection of methanol and other organics. *Mon. Not. R. Astron. Soc.* **265**, 1022–1026 (1993).
57. Clemett, S. J. & Maechling, C. R. Identification of complex aromatic molecules in individual interplanetary dust particles. *Science* **262**, 721–725 (1993).
58. Lewis, R. S., Ming, T., Wacker, J. F., Anders, E. & Steel, E. Interstellar diamonds in meteorites. *Nature* **326**, 160–162 (1987).
59. Nittler, L. R., Alexander, C. M. O'D., Gao, X., Walker, R. M. & Zinner, E. Stellar sapphires: the properties and origins of presolar Al₂O₃ in Meteorites. *Astrophys. J.* **483**, 475–495 (1997).
60. Bernatowicz, T. *et al.* Evidence for interstellar SiC in the Murray carbonaceous meteorite. *Nature* **330**, 728–730 (1987).
61. Zinner, E. Stellar nucleosynthesis and the isotopic composition of presolar grains from primitive meteorites. *Annu. Rev. Earth Planet. Sci.* **26**, 147–188 (1998).
62. Cronin, J. R., Pizzarello, S. & Frye, J. S. ¹³C NMR spectroscopy of the insoluble carbon of carbonaceous chondrites. *Geochim. Cosmochim. Acta* **51**, 299–303 (1987).
63. Flynn, G. J., Keller, L. P., Wirick, S. & Jacobsen, C. A preliminary assessment of the organic content of interplanetary dust particles. *Lunar Planet. Sci.* **34**, 1652 (2003).
64. Messenger, S. Identification of molecular-cloud material in interplanetary dust particles. *Nature* **404**, 968–971 (2000).
65. Brownlee, D. E. *et al.* Stardust: comet and interstellar dust sample return mission. *J. Geophys. Res.* **108**, SRD1–1 (2003).
66. Flynn, G. J., Keller, L. P., Feser, M., Wirick, S. & Jacobsen, C. The origin of organic matter in the solar system: evidence from the interplanetary dust particles. *Geophys. Cosmophys. Acta* **67**, 4791–4806 (2003).
67. Cody, G. D., Alexander, C. M. O'D. & Tera, F. Solid-state (¹H and ¹³C) nuclear magnetic resonance spectroscopy of insoluble organic residue in the Murchison meteorite: a self-consistent quantitative analysis. *Geophys. Cosmophys. Acta* **66**, 1851–1865 (2002).
68. Aléon, J., Robert, F., Chaussidon, M. & Marty, B. Nitrogen isotopic composition of macromolecular organic matter in interplanetary dust particles. *Geophys. Cosmophys. Acta* **67**, 3373–3783 (2003).

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to S.K. (kwok@asiaa.sinica.edu.tw).

Replication protein A interacts with AID to promote deamination of somatic hypermutation targets

Jayanta Chaudhuri, Chan Khuong & Frederick W. Alt

Howard Hughes Medical Institute, Children's Hospital, Center for Blood Research and Department of Genetics, Harvard University Medical School, Boston, Massachusetts 02115, USA

Activation-induced cytidine deaminase (AID) is a single-stranded (ss) DNA deaminase required for somatic hypermutation (SHM) and class switch recombination of immunoglobulin genes. Class switch recombination involves transcription through switch regions, which generates ssDNA within R loops. However, although transcription through immunoglobulin variable region exons is required for SHM, it does not generate stable ssDNA, which leaves the mechanism of AID targeting unresolved. Here we characterize the mechanism of AID targeting to *in-vitro*-transcribed substrates harbouring SHM motifs. We show that the targeting activity of AID is due to replication protein A (RPA), a ssDNA-binding protein involved in replication, recombination and repair. The 32-kDa subunit of RPA interacts specifically with AID from activated B cells in a manner that seems to be dependent on post-translational AID modification. Thus, our study implicates RPA as a novel factor involved in immunoglobulin diversification. We propose that B-cell-specific AID–RPA complexes preferentially bind to ssDNA of small transcription bubbles at SHM 'hotspots', leading to AID-mediated deamination and RPA-mediated recruitment of DNA repair proteins.

Three somatic DNA alteration events enable mammalian B lymphocytes to generate enormous antibody diversity. Antibodies are comprised of immunoglobulin heavy (IgH) and light (IgL) chains. Each chain has an amino-terminal variable region, involved in antigen binding, and a carboxy-terminal constant region, which, for IgH chains, determines downstream functions. Developing B cells use V(D)J recombination to assemble exons that encode the IgH and IgL variable (V) regions upstream of the corresponding constant region (C_H) exons¹. After transcription from a 5' V region exon promoter, primary transcripts are processed to form complete immunoglobulin (V and C) messenger RNAs. Newly generated B cells migrate to secondary lymphoid organs where, upon encountering antigens, they can be stimulated to undergo two additional immunoglobulin gene alterations: IgH class switch recombination (CSR) and SHM^{2–6}. Although they are distinct processes, both CSR and SHM absolutely require AID^{7,8}. How AID functions in these processes has been a subject of great interest and debate.

CSR allows activated B cells to exchange the IgH C_μ exons for an alternative set of C_H exons, permitting the expression of different antibody classes. CSR occurs via a recombination/deletion reaction within switch (S) regions that precede individual sets of C_H exons^{2,3,6}. Mammalian S regions are 1–12 kilobases (kb) in length and G-rich on the non-template strand². CSR is probably initiated by DNA double-strand breaks in two participating S regions and is completed by their fusion through general DNA repair⁶. In contrast, SHM introduces non-templated point mutations, with occasional insertions and deletions, at a very high rate into the V regions of immunoglobulin genes, ultimately allowing selection of cells that produce antibodies with increased affinity⁵. The RGYW (where R = purine nucleotides, Y = pyrimidine nucleotides, and W = A or T) DNA motif, and its inverse complement WRCY, form intrinsic SHM hotspots^{9,10}. A total of 50–60% of all SHMs are found around RGYW motifs, suggesting that this motif might target a 'mutator' complex¹⁰. In addition to AID dependency, CSR and SHM also require transcription through target V or S regions, which is thought to target initiation of these processes^{6,11,12}.

The role of AID has been a subject of debate. Some argue that AID functions as an RNA-editing enzyme³; however, there is no direct

evidence to support this hypothesis. In contrast, ectopically expressed AID associates with S regions in activated B cells¹³, and AID has cytidine deaminase activity on ssDNA^{14–17}. Genetic data led to the proposal of a model whereby deamination of deoxycytidines to deoxyuridines by AID would create dU·dG lesions that could be differentially processed to generate V region mutations or S region double strand breaks¹⁸. However, given AID specificity for ssDNA, the question arose as to how it could also function on double-stranded (ds) targets. Notably, transcription through mammalian S regions in the physiological, G-rich orientation generates R loops in which the transcript stably hybridizes to the template strand, thereby displacing the non-template strand as ssDNA^{19,20}. Correspondingly, synthetic purine-rich substrates lacking characteristic S region repeats served as effective AID substrates on *in vitro* transcription^{14,21} and, where examined, formed R loops *in vitro*^{19,22} and supported CSR *in vivo*²². However, although the ability of S regions to form R loops may provide, at least in part, a basis for the generation of an AID CSR substrate, V region exons are not G-rich and fail to form R loops or other known secondary structures, leaving a large degree of uncertainty surrounding the issue of how AID could be targeted during SHM.

SHMs are found in V region exons starting about 200 base pairs (bp) downstream of the promoter and extending through the exon for about 1–2 kb^{4,23}. In this regard, transcription and transcriptional regulatory sequences have a very important role in SHM^{24–26}. Of note, some non-variable region sequences also serve as targets for SHM in activated B cells, indicating a lack of absolute sequence specificity^{27–29}. However, the overall mutation load of SHM targets seems to correlate with SHM hotspot content³⁰. Together, these findings suggest that transcription is involved in generating SHM targets preferentially in sequences around RGYW motifs²⁵. On the basis of AID specificity^{14–17}, it is reasonable to speculate that such targets involve ssDNA. This general line of reasoning has led to the proposal that unknown factors cooperate with AID to effect SHM^{5,11}. As S regions also contain a high density of RGYW motifs, such factors might also operate with R loops to target CSR.

To search for hypothetical AID cofactors, we used an *in vitro* assay for activities that permit AID-mediated deamination of an actively

Table 1 DNA deamination activity in protein fractions from activated B cells of wild-type and *Aid*^{-/-} mice

Fraction	ssDNA deaminase activity (%)	Transcription-coupled deaminase activity (%) (RGYW substrate)
Nuclear extract	0.7/0.8	0.5/0.6
DEAE-cellulose	1.1/8.5	1.2/7.6
Glycerol gradient	0.8/9.2	1.1/1.2
Hydroxyapatite	1.2/9.8	1.2/0.9
Phosphocellulose	1.6/19.5	1.1/0.9
Anti-AID affinity	0.8/22.5	0.6/0.7

Values either side of the solidus are for *Aid*^{-/-} and wild type, respectively.

transcribed SHM substrate. On the basis of this approach, we identify RPA as a novel factor in immunoglobulin gene diversification.

DNA deamination of transcribed SHM substrates

Although AID is firmly established as a ssDNA deaminase, many questions remain regarding its *in vivo* function, including identification of putative cofactors or potential modifications. To address such issues, we devised an AID purification scheme (Table 1) from activated splenic B cells that generated AID of apparent electrophoretic homogeneity (Supplementary Fig. 1). AID from purified B cells (referred to hereafter as AID(Bcell)) exhibited robust ssDNA deaminase activity (Table 1) and deaminated dsDNA transcribed from T7 bacteriophage RNA polymerase that formed R loops upon *in vitro* transcription (Fig. 1d). However, AID(Bcell) did not

detectably deaminate a T7-polymerase-transcribed dsDNA model SHM substrate (referred to hereafter as RGYW substrate (Table 1)). The 540-bp RGYW substrate is comprised of 135 RGYW (and WRCY complement) motifs and lacks the ability to form detectable R loops upon *in vitro* transcription (data not shown).

We reasoned that a factor-dependent process exists that targets AID to transcribed RGYW-containing V genes *in vivo*, and that our *in-vitro*-transcription-coupled cytidine deamination assay on the RGYW substrate might be used to reveal such activities. Therefore, we assayed protein fractions generated en route to purification of AID(Bcell) for the ability to stimulate RGYW substrate deamination, and found that a crude AID(Bcell) preparation (DEAE fraction, Table 1) had significant deaminase activity on the transcribed RGYW substrate. Moreover, this activity was lost upon subsequent sedimentation through high-salt glycerol gradients (Table 1). We reasoned that the glycerol gradient fractions should contain the AID-complementing activity (that is, the complementing factor (CF)) and assayed each for the ability to reconstitute AID-mediated deamination. Indeed, we observed a sharp peak of CF activity, sedimenting at about 120–170 kDa (Fig. 1a). Gradient fractions containing CF activity were pooled and characterized further.

Both [³H]uracil release (Fig. 1b) and DNA cleavage cytidine deaminase assays (Fig. 1c) showed that CF, by itself, had no activity on either ssDNA or the transcribed RGYW substrate. The CF fraction also had no stimulatory effect on the ssDNA deaminase activity of AID (Fig. 1b). Moreover, the ability of CF to promote AID-mediated deamination on the RGYW substrate absolutely

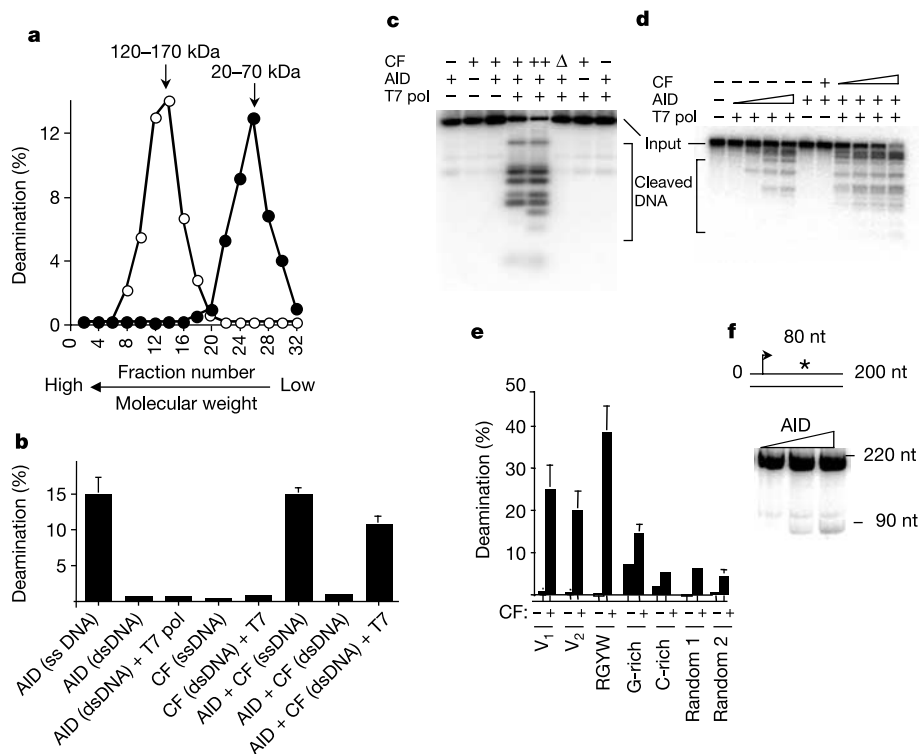


Figure 1 Transcription-coupled deamination on SHM substrates. **a**, Glycerol gradient fractions (2 μg) derived from activated B-cell extracts were assayed for ssDNA deaminase activity, without added purified AID (filled circles) or for dsDNA deaminase activity on transcribed RGYW substrate (open circles) in the presence of purified AID. **b**, **c**, CF activity (1 μg) on RGYW substrate (540 bp) was characterized by [³H]-release assay and gel cleavage assays, respectively. The 'Δ' represents heat-treated CF. **d**, CF activity on switch region (S_μ) DNA (3.2 kb) was analysed by DNA cleavage assay. **e**, Multiple DNA elements with varying RGYW content were tested as substrates in the transcription-coupled reaction. Substrates are indicated and described in more detail in the main text

and Methods. For cleavage assays, the RGYW DNA or S_μ DNA was used as hybridization probes. **f**, A 200-nucleotide DNA fragment that contained a single AGCT hotspot motif (asterisk), but having several deoxycytidine residues, was labelled at the 5'-end with ³²P and then subjected to AID-RPA-mediated deamination and DNA cleavage. The reaction products were analysed on a denaturing 8% polyacrylamide gel. DNA fragments of known molecular size were run as size markers in parallel lanes (not shown). For panels **b** and **e**, values represent the mean from three independent experiments, and error bars depict standard deviation from the mean.

depended on transcription. Thus, little or no deamination was observed when T7 RNA polymerase (Fig. 1b) or ribonucleotides (not shown) were omitted. Furthermore, CF activity was heat labile (Fig. 1c), suggesting a protein component. In addition, we found CF activity in nuclear extracts both from AID-deficient B cells and from non-lymphoid cell lines, indicating generalized expression (data not shown). Finally, we tested CF activity on DNA from a transcribed S region and found that, corresponding to the R-loop-forming ability, it displayed significant levels of deamination in the presence of purified AID alone. However, we found that CF markedly enhanced S region deamination compared to AID alone (Fig. 1d). We conclude that CF is a generally expressed factor with a protein component, and is capable of promoting AID-mediated, transcription-coupled DNA deamination *in vitro* on both SHM and S region substrates.

Several DNA sequences were tested as substrates in the CF/AID-mediated deamination assay (Fig. 1e, f, Supplementary Fig. 2a, and data not shown). The CF fraction stimulated AID-catalysed deamination on all sequences tested after transcription; however, the level of deamination achieved varied substantially among the substrates. Deamination of two different V genes rich in RGYW or WRCY hotspot motifs (approximately 25 and 20 motifs per 500 bp, respectively) was markedly stimulated (>25-fold) by CF above low baseline activity, with reaction efficiency approaching that of the RGYW substrate. In contrast, a G-rich random sequence (with approximately 10 SHM motifs in 1 kb) known to support CSR and R loops²² underwent significant levels of CF-independent deamination, whereas its inverse C-rich sequence did not, and deamination of both sequences was only modestly stimulated (2–6-fold) by CF. Furthermore, two randomly selected genomic DNA sequences (with approximately 5 and 6 SHM motifs per 500 and 400 bp, respectively) had low baseline activity and both showed only modest stimulation (2–5-fold) compared to the SHM-rich sequences (Fig. 1e). Finally, a substrate lacking RGYW motifs was not deaminated above background levels with this assay (Supplementary Fig. 2a). These results are consistent with the findings of transgenic studies that random sequences can serve as SHM substrates, with the frequency of mutations and, to some degree, specificity, being proportional to the number of hotspots they contain³⁰. To investigate more directly the role of the RGYW sequence, we tested a substrate that contained only one RGYW motif in a deamination/cleavage assay and found that deamination occurred predominantly at (or very near) the RGYW motif (Fig. 1f; see also Supplementary Fig. 2b). We conclude that CF activity promotes AID deamination of several *in-vitro*-transcribed DNA substrates assayed, and that both efficiency and targeting appear to be modulated by the RGYW motif.

To test whether the AID–CF complex targets deamination on both DNA strands of the transcribed RGYW substrate, reaction products from a cleavage assay were analysed by Southern blotting with probes specific to template or non-template strands. Only the non-template probe interacted with the cleaved DNA fragments, indicating that the template strand is not detectably deaminated (Fig. 2a). However, when transcriptional orientation was reversed using a T3 polymerase, similar overall deamination levels were observed via the [³H] release assay (Fig. 2b), indicating that both strands were capable of being deaminated if transcribed. Also, in agreement with our earlier results¹⁴, we observed that whereas AID predominantly targeted the non-template strand of *in-vitro*-transcribed S regions, the template strand was deaminated to a lesser degree (Fig. 2c). Although CF increased deamination of the transcribed S region substrate, it did not detectably enhance deamination of the template strand (Fig. 2c). We conclude that the AID–CF complex in our *in vitro* assays with T7 polymerase preferentially targets DNA deamination to the non-template strand.

Interaction of AID with complementing factor

The ability of recombinant AID to deaminate efficiently ssDNA *in vitro* argues against the requirement for specific cofactors for (at least) the ssDNA deaminase activity^{15–17}. However, the biochemical synergism between AID and CF in mediating deamination on the RGYW substrates prompted us to test whether AID and CF interact physically. First, we performed far-western analyses of different protein fractions using [³⁵S]AID(Bcell) (Fig. 3a) as a probe, and found several major interacting polypeptides in crude wild-type B-cell extracts (Fig. 3b). Consistent with the finding that AID dimerizes³¹, we observed a 26-kDa protein that appeared to be AID on the basis of its presence in wild-type extracts and purified AID preparations, but not in AID-deficient extracts. Another AID interactor, a 30–35-kDa protein, was present in both wild-type and AID-deficient extracts, and was highly enriched in fractions containing the highest CF activity (compare fractions 8–20 in Fig. 1a to those in Fig. 3b).

To confirm further the interaction between CF and AID, we partially purified CF activity from AID-deficient splenic B cells by glycerol gradient sedimentation, and incubated pooled CF fractions with AID immobilized on an anti-AID-antibody affinity column. Bound proteins were eluted with increasing NaCl concentration and assayed for CF activity (Fig. 3c). Whereas no CF activity was detected in the 0.25 M and 0.5 M NaCl eluates, the 0.75 M and 1.0 M fractions exhibited significant activity, indicating a tight association between AID and CF. Although we observed deaminase activity independent of added AID in the 1.0 M NaCl eluate, this was due to AID leaching off the affinity resin at this NaCl concentration (Fig. 3c, lower panel). Notably, the 30–35-kDa polypeptide that reacted with AID on far-western blots was found in the NaCl eluates containing the highest CF activity (Fig. 3c).

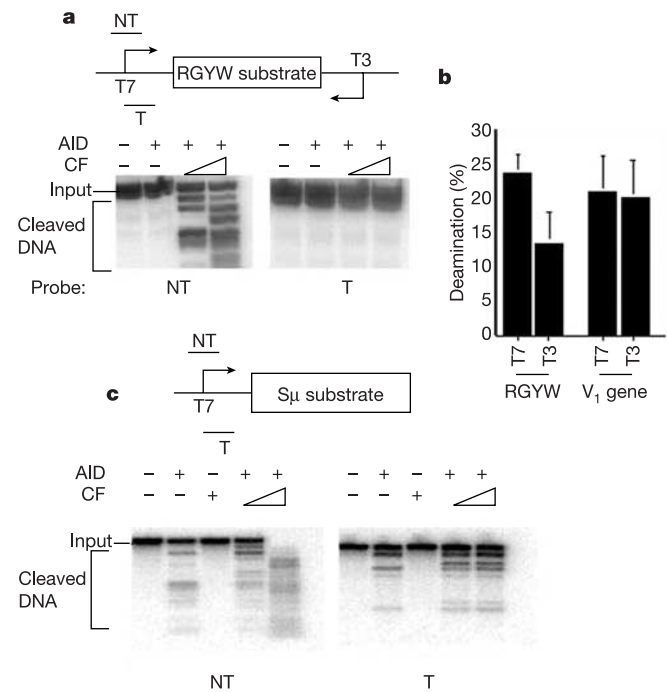


Figure 2 Strand-specific targeting of AID and CF. **a, c**, The RGYW DNA (540 bp) or the S_μ DNA (3.2 kb) was transcribed *in vitro* in the presence of AID (50 ng) and CF (0.5 μg) as indicated and then treated with UNG and hot alkali. The reaction products were divided into two aliquots and then subjected to Southern hybridization using oligonucleotide probes derived from the template (T) or the non-template (NT) strand. **b**, The RGYW or the V gene was transcribed in either orientation by T7 or T3 RNA polymerases, and deamination in the presence of both AID and CF was measured by the [³H]-release assay. Error bars represent the standard deviation from the mean (of three independent experiments).

These results suggest that a 30–35-kDa protein is a component of CF. However, the sedimentation of CF as a 120–170-kDa protein in glycerol gradients (Fig. 1a) indicates that the 30–35-kDa polypeptide is part of a larger protein complex. In this context, lack of interaction between CF and AID in the glycerol gradient probably reflects dissociation of AID from the complex owing to the hydrostatic force of sedimentation under high salt concentration.

RPA is the complementing factor

On the basis of the reactivity of AID with the 30–35-kDa interactor on far-western blots, we attempted to clone the complementary DNA encoding this protein by screening a lambda phage expression library using ^{35}S -labelled AID(Bcell) as a probe. Four clones, which interacted with AID but not with a nonspecific control probe (Fig. 3a, *Aid*^{-/-} probe), were identified and the cDNAs sequenced. All encoded the 32-kDa subunit (RPA32) of RPA, a protein of a size similar to that of the AID interactor. RPA is an abundant ssDNA-binding protein, originally identified as a protein essential for SV40 DNA replication^{32–34}, and subsequently demonstrated to have roles in cellular DNA replication, recombination and repair³⁵. It is a heterotrimeric protein consisting of 70 kDa (RPA70), 32 kDa (RPA32) and 14 kDa (RPA14) polypeptides, with RPA70 serving as the major DNA-binding subunit. To confirm whether the CF fraction contained RPA, we used western blotting with antibodies specific to RPA70 and RPA32, and found that both subunits were highly enriched in CF-containing gradient fractions (Fig. 4a). Although anti-RPA14 antibodies are not readily available, the three RPA subunits associate together tightly³⁵. In addition, the approximate molecular weight of the CF complex, based on glycerol gradients, is appropriate for RPA.

Immunoprecipitations provided additional support for an interaction between RPA and AID, as RPA immunoprecipitates from wild-type, but not AID-deficient, B cells contained AID activity (Fig. 4b, upper panel). In addition, anti-AID antibodies specifically immunoprecipitated RPA from wild-type, but not AID-deficient,

B-cell extracts (Fig. 4b, lower panel). Treatment of AID and RPA with DNase demonstrated that the interaction is not indirect owing to the ssDNA-binding property of these proteins (Supplementary Fig. 3). These observations confirmed that AID interacts directly with RPA32. To demonstrate unequivocally that RPA is required for AID-catalysed deamination of the RGYW motif, we tested bacterially expressed recombinant human RPA (rRPA) containing all three subunits for the ability to replace CF, and found that it functioned quite comparably (Fig. 4c). Finally, we observed that other ssDNA-binding proteins (that is, bacterial SSB and bacteriophage T4 gp32) were not viable substitutes for RPA in AID-catalysed deamination (Fig. 4d), indicating specificity for RPA in the reaction.

Coordinate binding of RPA and AID to transcribed DNA

RPA binds ssDNA with very high affinity ($\sim 10^9 \text{ M}^{-1}$), but has low affinity for dsDNA³⁵. Similarly, AID binds ssDNA with much higher affinity than dsDNA¹⁵. However, we predicted that the novel activity of the two proteins together might result in a potential interaction with transcribed dsDNA of the RGYW substrate that might be detected by gel shifts. As expected, AID or RPA alone did not bind to the transcribed RGYW substrate, and together they did not bind to the substrate in the absence of transcription (Fig. 5a). However, addition of both proteins together to the transcribed substrate unmasks significant DNA-binding activity in the form of two

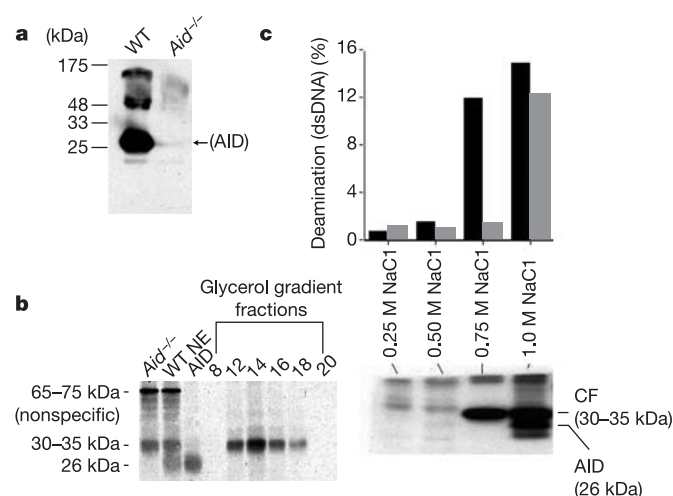


Figure 3 AID interacts with CF. **a**, [^{35}S]AID(Bcell) (50 ng) purified from wild-type (WT) B cells or similar fractions from AID-deficient (*Aid*^{-/-}) cells were analysed by SDS gels followed by autoradiography. **b**, Glycerol gradient fractions (Fig. 1a), purified AID and protein preparations as marked (1–10 μg) were subjected to far-western analysis using [^{35}S]AID(Bcell) as a probe. The 65–75-kDa protein also interacted with the protein preparation purified from *Aid*^{-/-} B cells; therefore, it was considered nonspecific. NE, nuclear extract. **c**, Partially purified nuclear extract obtained from *Aid*^{-/-} B cells was fractionated through an immobilized AID-affinity column. Bound proteins were eluted with increasing concentrations of NaCl and assayed for deamination reaction in the presence (black bars) or absence (grey bars) of added AID on transcribed RGYW substrate (top panel), or for the ability to interact with AID on far-western blots (bottom panel).

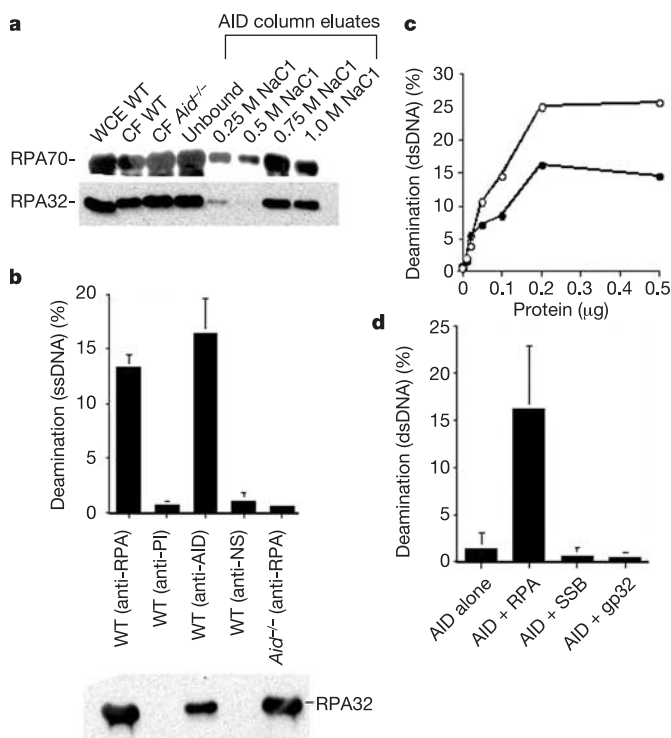


Figure 4 The complementing factor is RPA. **a**, Different CF fractions (0.25 μg), including those eluting from the AID-affinity column, were analysed by western blotting using RPA32 and RPA70 antibodies. WCE, whole-cell extract. **b**, Partially purified extracts from wild-type or *Aid*^{-/-} B cells were subjected to immunoprecipitation using AID or RPA antibodies. The immunoprecipitates were assayed for ssDNA deaminase activity (top panel) or by western blotting using RPA32 antibodies (bottom panel). PI and NS represent two nonspecific antibodies used as controls. **c**, Purified recombinant human RPA (open circles) and CF (filled circles) were assayed for AID-catalysed DNA deamination on RGYW substrate. **d**, Two ssDNA-binding proteins, *E. coli* SSB and T4 bacteriophage gp32, were assayed for their ability to replace RPA in the deamination reaction. In panels **b**, **d**, values represent the mean of three independent experiments, and error bars depict standard deviation from the mean.

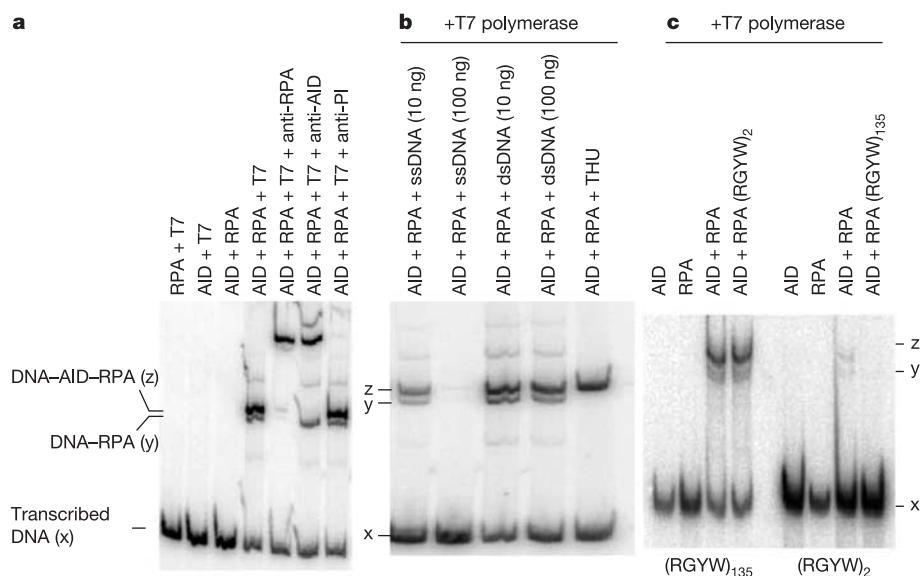


Figure 5 Coordinate binding of AID and RPA to transcribed SHM substrate. **a**, **b**, 32 P-labelled RGYW DNA (180 bp) was used as a substrate in gel shift experiments using purified RPA (0.1 μ g), AID (0.05 μ g) and other components as indicated. THU,

tetrahydropyrimidine (25 μ M). **c**, Competition experiments were carried out in the presence of a tenfold excess of the indicated competitor. Unless otherwise indicated, the RGYW DNA substrate contained 135 repeats of the hotspot motif.

distinct DNA–protein complexes (Fig. 5a), the formation of which could be inhibited by excess, nonspecific ssDNA but not equivalent amounts of dsDNA (Fig. 5b). In addition, a substrate containing only two RGYW motifs neither bound the AID–RPA complex efficiently nor competed for binding to the RGYW substrate, indicating that binding is, at least in large part, dependent on RGYW (Fig. 5c).

Out of the two protein–RGYW DNA complexes that were observed, the larger-molecular-weight complex was supershifted by both AID- and RPA-specific antibodies, indicating that it contained both AID and RPA. However, the lower-molecular-weight complex probably contained only RPA, as it was supershifted by anti-RPA but not anti-AID antibodies (Fig. 5a). We suggest that the DNA–RPA complex probably derives from the DNA–RPA–AID complex via dissociation of AID, as the former was not observed in the absence of AID. In this regard, the DNA–RPA complex was not observed when the deamination reaction was inhibited by the cytidine deaminase inhibitor tetrahydropyrimidine (THU; Fig. 5b), even though formation of the DNA–RPA–AID complex was not impaired. Together, these results suggest that AID and RPA are both required for formation of a complex on transcribed DNA, and that AID-catalysed deamination of the substrate leads to release of AID, leaving the RPA still bound.

B-cell specificity of the AID–RPA interaction

In earlier studies, we observed that both AID(Bcell) and ectopically expressed AID from 293 cells (AID(293)) could deaminate ssDNA with similar efficiency¹⁴. However, purified AID(293), although having similar ssDNA activity to AID(Bcell) (Fig. 6d), had little activity with respect to the deamination of or binding to the RGYW substrate with added RPA (Fig. 6a and data not shown). This altered activity of AID(293) was not due to a nonspecific inhibitor of transcription-coupled deamination in 293 extracts, as addition of AID(Bcell) restored deamination (data not shown). Therefore, we considered the possibility that AID(293) does not interact with RPA. Indeed, immunoprecipitated RPA from 293 cells failed to bring down AID activity (Fig. 6b), and immunoprecipitated AID(293) failed to bring down RPA (Fig. 6c). There could be several explanations for these findings, including B-cell-specific, AID-associated factors or post-translational AID modifications. In this

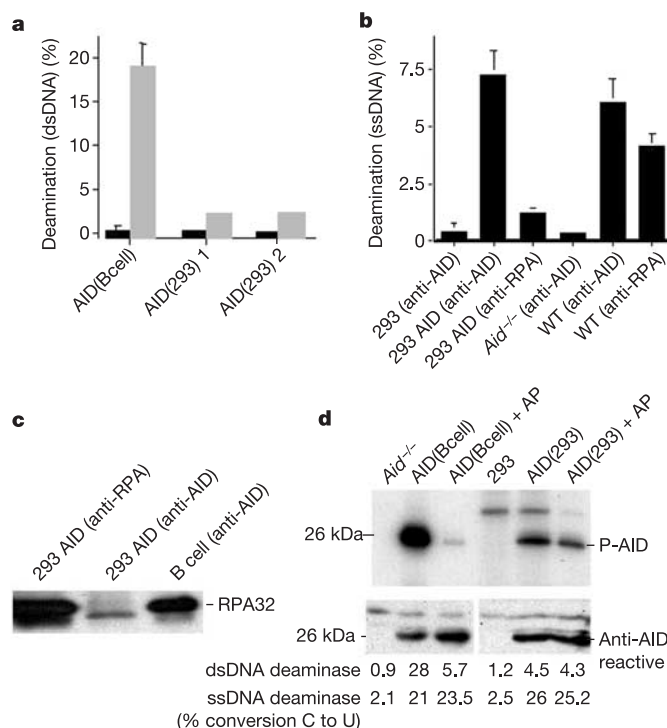


Figure 6 B-cell specificity of the AID–RPA complex. **a**, AID(Bcell) or two independent preparations of AID(293), all with similar ssDNA deamination activity, were assayed for dsDNA deamination in the presence (grey bars) or absence (black bars) of RPA. **b**, **c**, Extracts from 293 cells expressing AID or from activated splenic B cells (WT or AID^{−/−}) were immunoprecipitated using AID or RPA antibodies. The immunoprecipitates were analysed for the presence of AID by ssDNA deaminase assay (**b**) or RPA by western blotting (**c**) using RPA32 antibodies. **d**, AID was purified from 32 P-labelled B cells or ectopically expressing 293 cells, treated with alkaline phosphatase (AP) and analysed by SDS–polyacrylamide gels followed by autoradiography (top panel) or by western blotting (bottom panel) using anti-AID antibodies. The proteins were also assayed for single-stranded- or RPA-mediated transcription-coupled dsDNA deaminase activity. In panels **a** and **b** values represent the mean of three independent experiments and error bars represent standard deviation from the means.

context, during the course of AID purification from B cells, we observed loss of dsDNA deaminase activity on the transcribed RGYW substrate, without loss of ssDNA deaminase activity, and this loss could be avoided by adding phosphatase inhibitors (data not shown). To search for potential AID phosphorylation, we purified AID from ^{32}P -labelled activated B cells or ectopically expressing 293 cells, and observed that AID(Bcell) had a substantially higher level of incorporated ^{32}P than did AID(293) for a comparable amount of protein and ssDNA activity (Fig. 6d). In addition, we tested the sensitivity of purified AID(Bcell) to alkaline phosphatase and found that phosphatase treatment substantially removed the ^{32}P from AID(Bcell), impaired the interaction with RPA (Supplementary Fig. 4), and markedly reduced deaminase activity on the transcribed RGYW substrate, but not on ssDNA substrates (Fig. 6d; see also Supplementary Fig. 5). Furthermore, most of the ^{32}P in AID(Bcell) appeared to be more sensitive to phosphatase than that of AID(293), suggesting the presence of more than one phosphorylated residue with potentially distinct functional significance (Fig. 6d). Similar phosphatase treatment of RPA had no effect on its activity in this assay (Supplementary Fig. 5). Together, these findings indicate that post-translational modifications of AID in activated B cells can modulate deaminase activity on transcribed SHM substrates by influencing RPA–AID interactions.

Discussion

We identify RPA, a ubiquitous ssDNA-binding protein, as the factor that targets AID to SHM motifs (for overall model, see Supplementary Fig. 7). Stabilization of ssDNA³⁵ and deamination of deoxycytidines at RGYW motifs^{21,36} are respective functions of RPA and AID. However, coordinate activity of these proteins is required to access *in-vitro*-transcribed SHM substrates. As RPA can bind single-stranded transcription bubbles in dsDNA as small as four nucleotides³⁷, we propose that the RPA–AID complex binds to and stabilizes ssDNA at small, transient transcription bubbles, with a preference for RGYW motifs. To our knowledge, RPA has not previously been found to bind to transcription bubbles; thus our current findings demonstrate a novel activity for RPA. After deamination, AID can be released from transcribed SHM substrates, leaving RPA DNA-bound. Such bound RPA might serve as a platform for entry of DNA repair proteins (as is the case in nucleotide excision repair³⁵), which have been proposed to lead to SHM^{5,18}. In the latter context, RPA has been implicated in DNA repair pathways thought to act downstream of AID deamination in SHM, including mismatch repair and, most notably, base excision repair via its interaction with UNG³⁸.

It is possible that AID–RPA complexes have a function in CSR by means of recruiting factors downstream of AID-catalysed DNA deamination. In addition, whereas AID may directly deaminate ssDNA in R loops during CSR, AID–RPA complexes might target S-region RGYW motifs to provide an additional CSR-targeting mechanism^{22,39}. In this regard, we have used chromatin immunoprecipitation assays to determine whether endogenous AID and RPA bind to S regions in an activation-dependent fashion. Notably, endogenous AID, similar to ectopically expressed AID¹³, binds in a highly specific manner to S regions in direct correlation to their activation for CSR, providing strong, direct support of a DNA deamination function of AID *in vivo* (Supplementary Fig. 6). On the other hand, RPA binds to several tested sequences, including S regions, in an activation- and AID-independent fashion in activated, highly proliferating B cells, consistent with the general role of RPA in DNA replication and repair. Of note, binding of RPA was reproducibly enriched to appropriate S regions in an activation- and AID-dependent fashion (Supplementary Fig. 6), supporting a potential role for the AID–RPA complex in CSR.

Analysis of the immunoglobulin gene mutation profile in B cells suggests that both DNA strands are targets of AID⁴⁰. Yet, with our

bacteriophage promoter/polymerase system we observed deamination largely on the non-template strand. We propose that exclusion of the template strand from deamination in this study and others^{21,41} may reflect the fact that the prokaryotic polymerase–DNA complex exposes only the non-template strand to modifying enzymes⁴². However, the eukaryotic RNA pol II–DNA complex is structurally distinct and can be modified by binding various transcription factors⁴³. Thus, RNA pol II, by itself or in association with other proteins, may facilitate binding of the AID–RPA complex to both strands. Alternatively, additional AID-interacting factors may be required to target the template strand. Also, anti-sense transcripts have been observed in the IgH locus⁴⁴, providing another conceivable mechanism for accessing the other DNA strand. Notably, both AID and RPA interact with RNA pol II holoenzyme^{13,45}; thus, in accord with one model for SHM²⁵, these proteins might travel with polymerase during elongation.

In B cells, AID activity is restricted to immunoglobulin genes and a small subset of non-immunoglobulin genes. However, overexpression of AID in non-lymphoid mammalian cells or bacteria more generally mutates transcribed genes^{41,46,47}, albeit at a lower rate, leading to the prediction that tightly regulated AID activity along with specific B-cell factors and *cis*-elements may be required for high-level SHM at V regions⁴⁷. In this context, the AID–RPA interaction is B-cell-specific *in vivo*, and the ability of AID to target SHM substrates *in vitro* seems to be, at least in part, dependent on AID phosphorylation. Therefore, B-cell-specific, post-translational AID modifications probably influence the ability to interact with RPA and (potentially) other factors, and thereby can modulate target specificity (Supplementary Fig. 7). Such a mechanism could explain why overexpression of AID in fibroblasts or bacteria leads primarily to transition mutations at G–C base pairs^{18,46,47}, as opposed to the transition and transversion mutations that are observed during SHM in B cells. Thus, whereas overexpressed AID in non-lymphoid cells may catalyse low-level, nonspecific deamination, uncoupling the AID–RPA interaction could abort UNG recruitment and the proposed downstream activation of transversion mutations, leaving G–U mismatches to be resolved only through DNA replication, with the resulting predominance of transitions^{48,49}. We conclude that the biochemical functions of the AID–RPA interaction provide a mechanistic explanation for aspects of the B-cell specificity of SHM. In this context, these functions might also be involved in translocation or mutation of oncogenes, which is often observed in mature B cell lymphomas. Finally, these functions strongly support the DNA deamination model of AID function and mechanistically link the SHM requirements of transcription, AID and RGYW motifs. □

Methods

DNA substrates

The RGYW somatic hypermutation substrate has been described previously¹⁴. The 180-bp fragment was trimerized to generate the 540-bp DNA used in the deamination assays. The B1-8 J558 variable region DNA (V₁) and the FO6 J558 (V₂) fragments were obtained by polymerase chain reaction (PCR) from mouse germinal centre cells. The fragments were cloned into pBluescript and labelled with [^3H]dCTP as described previously¹⁴. We also used PCR-amplified 400–500-bp intragenic genomic sequences, which were not G-rich and contained few RGYW motifs, as substrates in the DNA deamination assays (see main text for a more detailed description of RGYW motif content of each sequence).

Deamination assays

The [^3H] release and gel cleavage assays were carried out as described previously¹⁴. For the cleavage assays, the test fragment with the T7 promoter was either PCR-amplified (for RGYW substrate) or excised from the plasmid (for switch DNA) with appropriate restriction enzymes.

Protein purification

For purification of AID, mouse splenic B cells were purified and stimulated for 96 h as described¹⁴, and nuclear extracts were prepared by extracting nuclear proteins with 0.6 M NaCl. The nuclear extract (that had no ssDNA or dsDNA deaminase activity owing to nonspecific inhibition by nucleic acids) was incubated with protein G-sepharose to remove contaminating immunoglobulins, and subsequently fractionated via DEAE-cellulose,

glycerol gradient sedimentation (100 mM NaCl) and hydroxyapatite columns¹⁴. The protein was further purified via chromatography through heparin sepharose (binding at 50 mM NaCl and elution at 125 mM), phosphocellulose (loading at 65 mM NaCl and elution with a linear NaCl gradient between 100 mM and 600 mM) and finally incubated with an affinity column in which anti-AID antibodies were immobilized on protein A-agarose beads. Bound AID was eluted with 1 M glycine pH 3.5, immediately diluted with 1 M Tris-Cl and dialysed against 55% glycerol buffer containing 20 mM Tris-Cl, pH 7.5, 100 mM NaCl, 10 μ M ZnCl₂, 1 mM dithiothreitol. The protein was stored at -20 °C. AID overexpressed in 293 cells was purified by the same procedure, without the initial protein G-sepharose step. For purification of [³⁵S]AID, activated B cells were labelled continuously with [³⁵S]methionine in complete DMEM media for 90 h, then starved in methionine-free media for 1 h, followed by the addition of [³⁵S]methionine. Cells were collected 5 h later and extracts prepared as described above.

Recombinant three-subunit human RPA was purified from *Escherichia coli* (provided by B. Stillman) as described previously⁵⁰.

Immunoprecipitation

Immunoprecipitation experiments were performed on the DEAE-cellulose purified protein fractions, using standard techniques as described previously¹⁴. Anti-AID antibodies have been described¹⁴ and RPA32 and RPA70 antibodies were obtained from Santa Cruz and Oncogene Sciences.

Far-western analysis

Proteins were fractionated on SDS-polyacrylamide gels and transferred to PVDF membranes. The blot was denatured in 6 M guanidine-HCl solution and then renatured by serial dilutions with PBS. The renatured blot was incubated with [³⁵S]AID for 24 h, washed and exposed for 24 h on a phosphorimager screen.

To clone the AID interactor, 5 × 10⁵ clones from a lambda zap expression library (Stratagene) were induced with IPTG, the expressed proteins immobilized on nitrocellulose membranes, and then subjected to denaturation/renaturation as described above. The membranes were incubated with [³⁵S]AID, washed and then exposed on films for 96–144 h. Positive clones were purified via four rounds of screening, the DNA was excised according to the manufacturer's protocol and nucleotide sequence determined by standard methods.

Gel shift assays

The 180-bp RGYW substrate was end-labelled with ³²P using T4 polynucleotide kinase. The labelled substrate was transcribed with T7 RNA polymerase in the presence of RPA and AID for 30 min at room temperature. Transcription was stopped by adding heparin, and after an additional incubation for 30 min, the reaction was analysed by electrophoresis on a 5% non-denaturing polyacrylamide gel.

Received 8 June; accepted 12 July 2004; doi:10.1038/nature02821.

Published online 25 July 2004.

- Jung, D. & Alt, F. W. Unraveling V(D)J recombination: Insights into gene regulation. *Cell* **116**, 299–311 (2004).
- Manis, J. P., Tian, M. & Alt, F. W. Mechanism and control of class-switch recombination. *Trends Immunol.* **23**, 31–39 (2002).
- Honjo, T., Kinoshita, K. & Muramatsu, M. Molecular mechanism of class switch recombination: linkage with somatic hypermutation. *Annu. Rev. Immunol.* **20**, 165–196 (2002).
- Papavasiliou, F. N. & Schatz, D. G. Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. *Cell* **109**, S35–S44 (2002).
- Li, Z., Woo, C. J., Iglesias-Ussel, M. D., Ronai, D. & Scharff, M. D. The generation of antibody diversity through somatic hypermutation and class switch recombination. *Genes Dev.* **18**, 1–11 (2004).
- Chaudhuri, J. & Alt, F. W. Class switch recombination: interplay of transcription, DNA deamination and DNA repair. *Nature Rev. Immunol.* **4**, 541–552 (2004).
- Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563 (2000).
- Revy, P. *et al.* Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell* **102**, 565–575 (2000).
- Betz, A. G., Neuberger, M. S. & Milstein, C. Discriminating intrinsic and antigen-selected mutational hotspots in immunoglobulin V genes. *Immunol. Today* **14**, 405–411 (1993).
- Neuberger, M. S. *et al.* Monitoring and interpreting the intrinsic features of somatic hypermutation. *Immunol. Rev.* **162**, 107–116 (1998).
- Storb, U. & Stavnezer, J. Immunoglobulin genes: generating diversity with AID and UNG. *Curr. Biol.* **12**, 725–727 (2002).
- Yu, K. & Lieber, M. R. Nucleic acid structures and enzymes in the immunoglobulin class switch recombination mechanism. *DNA Repair* **2**, 1163–1174 (2003).
- Nambu, Y. *et al.* Transcription-coupled events associating with immunoglobulin switch region chromatin. *Science* **302**, 2137–2140 (2003).
- Chaudhuri, J. *et al.* Transcription-targeted DNA deamination by the AID antibody diversification enzyme. *Nature* **422**, 726–730 (2003).
- Dickerson, S. K., Market, E., Besmer, E. & Papavasiliou, F. N. AID mediates hypermutation by deaminating single stranded DNA. *J. Exp. Med.* **197**, 1291–1296 (2003).
- Bransteitter, R., Pham, P., Scharff, M. D. & Goodman, M. F. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. *Proc. Natl Acad. Sci. USA* **100**, 4102–4107 (2003).
- Sohail, A., Klapacz, J., Samaranyake, M., Ullah, A. & Bhagwat, A. S. Human activation-induced cytidine deaminase causes transcription-dependent, strand-biased C to U deaminations. *Nucleic Acids Res.* **31**, 2990–2994 (2003).
- Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* **418**, 99–103 (2002).
- Tian, M. & Alt, F. W. Transcription-induced cleavage of immunoglobulin switch regions by nucleotide

- excision repair nucleases *in vitro*. *J. Biol. Chem.* **275**, 24163–24172 (2000).
- Yu, K., Chedin, F., Hsieh, C. L., Wilson, T. E. & Lieber, M. R. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nature Immunol.* **4**, 442–451 (2003).
- Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424**, 103–107 (2003).
- Shinkura, R. *et al.* The influence of transcriptional orientation on endogenous switch region function. *Nature Immunol.* **4**, 435–441 (2003).
- Storb, U. *et al.* Cis-acting sequences that affect somatic hypermutation of Ig genes. *Immunol. Rev.* **162**, 153–160 (1998).
- Betz, A. G. *et al.* Elements regulating somatic hypermutation of an immunoglobulin kappa gene: critical role for the intron enhancer/matrix attachment region. *Cell* **77**, 239–248 (1994).
- Peters, A. & Storb, U. Somatic hypermutation of immunoglobulin genes is linked to transcription initiation. *Immunol.* **4**, 57–65 (1996).
- Fukita, Y., Jacobs, H. & Rajewsky, K. Somatic hypermutation in the heavy chain locus correlates with transcription. *Immunol.* **9**, 105–114 (1998).
- Yelamos, J. *et al.* Targeting of non-Ig sequences in place of the V segment by somatic hypermutation. *Nature* **376**, 225–229 (1995).
- Pasqualucci, L. *et al.* BCL-6 mutations in normal germinal center B cells: evidence of somatic hypermutation acting outside Ig loci. *Proc. Natl Acad. Sci. USA* **95**, 11816–11821 (1998).
- Shen, H. M., Peters, A., Baron, B., Zhu, X. & Storb, U. Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of Ig genes. *Science* **280**, 1750–1752 (1998).
- Michael, N. *et al.* Effects of sequence and structure on the hypermutability of immunoglobulin genes. *Immunol.* **16**, 123–134 (2002).
- Ta, V. T. *et al.* AID mutant analyses indicate requirement for class-switch-specific cofactors. *Nature Immunol.* **4**, 843–848 (2003).
- Wobbe, C. R. *et al.* Replication of simian virus 40 origin-containing DNA *in vitro* with purified proteins. *Proc. Natl Acad. Sci. USA* **84**, 1834–1838 (1987).
- Wold, M. S. & Kelly, T. Purification and characterization of replication protein A, a cellular protein required for *in vitro* replication of simian virus 40 DNA. *Proc. Natl Acad. Sci. USA* **85**, 2523–2527 (1988).
- Fairman, M. P. & Stillman, B. Cellular factors required for multiple stages of SV40 DNA replication *in vitro*. *EMBO J.* **7**, 1211–1218 (1988).
- Wold, M. S. Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu. Rev. Biochem.* **66**, 61–92 (1997).
- Yu, K., Huang, F. T. & Lieber, M. R. DNA substrate length and surrounding sequence affect the activation-induced deaminase activity at cytidine. *J. Biol. Chem.* **279**, 6496–6500 (2004).
- Matsunaga, T., Park, C. H., Bessho, T., Mu, D. & Sancar, A. Replication protein A confers structure-specific endonuclease activities to the XPF-ERCC1 and XPG subunits of human DNA repair excision nuclease. *J. Biol. Chem.* **271**, 11047–11050 (1996).
- Nagelhus, T. A. *et al.* A sequence in the N-terminal region of human uracil-DNA glycosylase with homology to XPA interacts with the C-terminal part of the 34-kDa subunit of replication protein A. *J. Biol. Chem.* **272**, 6561–6566 (1997).
- Tashiro, J., Kinoshita, K. & Honjo, T. Palindromic but not G-rich sequences are targets of class switch recombination. *Int. Immunol.* **13**, 495–505 (2001).
- Milstein, C., Neuberger, M. S. & Staden, R. Both DNA strands of antibody genes are hypermutation targets. *Proc. Natl Acad. Sci. USA* **95**, 8791–8794 (1998).
- Ramiro, A. R., Stavropoulos, P., Jankovic, M. & Nussenzweig, M. C. Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. *Nature Immunol.* **4**, 452–456 (2003).
- Artsimovitch, I. & Landick, R. The transcriptional regulator RfaH stimulates RNA chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand. *Cell* **109**, 193–203 (2002).
- Steitz, T. A. The structural basis of the transition from initiation to elongation phases of transcription, as well as translocation and strand separation, by T7 RNA polymerase. *Curr. Opin. Struct. Biol.* **14**, 4–9 (2004).
- Bolland, D. J. *et al.* Antisense intergenic transcription in V(D)J recombination. *Nature Immunol.* **5**, 630–637 (2004).
- Maldonado, E. *et al.* A human RNA polymerase II complex associated with SRB and DNA-repair proteins. *Nature* **381**, 86–89 (1996).
- Yoshikawa, K. *et al.* AID enzyme-induced hypermutation in an actively transcribed gene in fibroblasts. *Science* **296**, 2033–2036 (2002).
- Martin, A. & Scharff, M. D. Somatic hypermutation of the AID transgene in B and non-B cells. *Proc. Natl Acad. Sci. USA* **99**, 12304–12308 (2002).
- Di Noia, J. & Neuberger, M. S. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature* **419**, 43–48 (2002).
- Rada, C. *et al.* Immunoglobulin isotype switching is inhibited and somatic hypermutation perturbed in UNG-deficient mice. *Curr. Biol.* **12**, 1748–1755 (2002).
- Henricksen, L. A., Umbricht, C. B. & Wold, M. S. Recombinant replication protein A: expression, complex formation, and functional characterization. *J. Biol. Chem.* **269**, 11121–11132 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Honjo for providing AID-deficient mice and B. Stillman for providing bacterial strains expressing recombinant AID. We also thank M. Tian, A. Zarrin, J. Manis and C. Bassing for critical reading of the manuscript, D. Chowdhury for advice on the ChIP experiments, and B. Stillman and S. Elledge for discussions. This work was supported by an NIH grant to F.W.A. F.W.A. is an Investigator of the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to F.W.A. (alt@enders.tch.harvard.edu).

The Lyman- α glow of gas falling into the dark matter halo of a $z = 3$ galaxy

Michael Weidinger^{1,2}, Palle Møller¹ & Johan Peter Uldall Fynbo^{2,3}

¹European Southern Observatory, Karl-Schwarzschild-Straße 2, D-85748 Garching bei München, Germany

²Institute of Physics and Astronomy, University of Aarhus, Ny Munkegade, DK-8000 Århus C, Denmark

³Astronomical Observatory, University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen Ø, Denmark

Quasars are the visible signatures of gas falling into the deep potential well of super-massive black holes in the centres of distant galaxies. It has been suggested¹ that quasars are formed when two massive galaxies collide and merge, leading to the prediction that quasars should be found in the centres of regions of largest overdensity in the early Universe. In dark matter (DM)-dominated models of the early Universe, massive DM halos are predicted to attract the surrounding gas, which falls towards their centres. The neutral gas is not detectable in emission by itself, but gas falling into the ionizing cone of such a quasar will glow in the Lyman- α line of hydrogen, effectively imaging the DM halo². Here we present a Ly α image of a DM halo at redshift $z = 3$, along with a two-dimensional spectrum of the gaseous halo. Our observations are best understood in the context of the standard model for DM haloes³; we infer a mass of $(2 - 7) \times 10^{12}$ solar masses ($M_{\odot}$) for the halo.

Using radiative transfer calculations, Haiman and Rees² predicted that gas falling into a DM halo between redshifts 3 and 8 that was harbouring a quasar should be detectable in Ly α -emission at flux levels accessible to present-day telescopes, owing to the reprocessing of quasar ultraviolet photons. Recently, Barkana and Loeb⁴ found absorption features in quasar spectra, which they interpreted as a signature of neutral hydrogen (H I) falling into the DM halos surrounding two quasars. Such absorption features can be used to study the gas in one dimension (1D) along the line of sight,

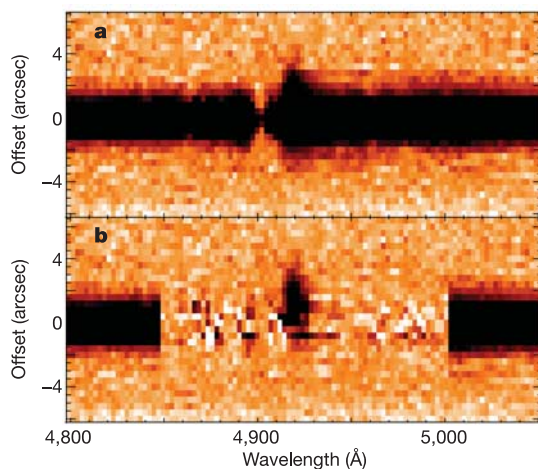


Figure 1 Two-dimensional spectrum of Q1205-30 and the extended Ly α emission. The wavelength increases along the abscissa, the position on the sky changes along the ordinate. The spectroscopy was performed with a slit position angle of 7.9° east of north (see Fig. 2a). **a**, The 2D quasar spectrum. The extended Ly α emission is faintly visible at $4,920 \text{ \AA}$. **b**, The quasar spectrum has been subtracted between the wavelengths $4,850\text{--}5,000 \text{ \AA}$, clearly revealing the underlying extended emission.

whereas detection of extended Ly α -emission can be used in a two-dimensional (2D) study, providing more constraints on the interplay between gas, quasar radiation and the DM halo.

In a deep Ly α narrow-band image we detected asymmetric extended emission northeast of the $z = 3$ radio-quiet quasar Q1205-30, making this quasar a good candidate for a study of its DM halo. The extended Ly α emission was first thought to be related to a foreground absorber⁵, but deep follow-up spectroscopy obtained with the FORS1 instrument on the ESO Very Large Telescope allowed us to measure precise redshifts of quasar, absorber and extended emission, clearly linking the extended emission to the quasar, not the absorber. The 2D spectrum is presented in Fig. 1 (details of the data reduction will be presented elsewhere). In the 2D spectrum the wavelength increases along the abscissa, and the position on the sky changes along the ordinate. In order to reveal the underlying extended emission we have optimally extracted the 2D quasar spectrum by fitting and removing the quasar spectral point spread function⁶. Spectroscopic detection of kinematically resolved extended Ly α emission is not uncommon^{7,8}, but it does not provide sufficient information to distinguish between scenarios of—for example—nearly edge-on disks and gaseous halos. For this, morphological information is needed. In Fig. 2a we present a $10 \times 10 \text{ arcsec}^2$ narrow-band image of the extended emission.

A quasar emits its radiation in an ionizing cone with an opening angle of ψ , while we observe the whole system under an inclination angle of θ (see Fig. 3). The surface brightness at each point in the vicinity of the quasar is calculated by integration of the volume

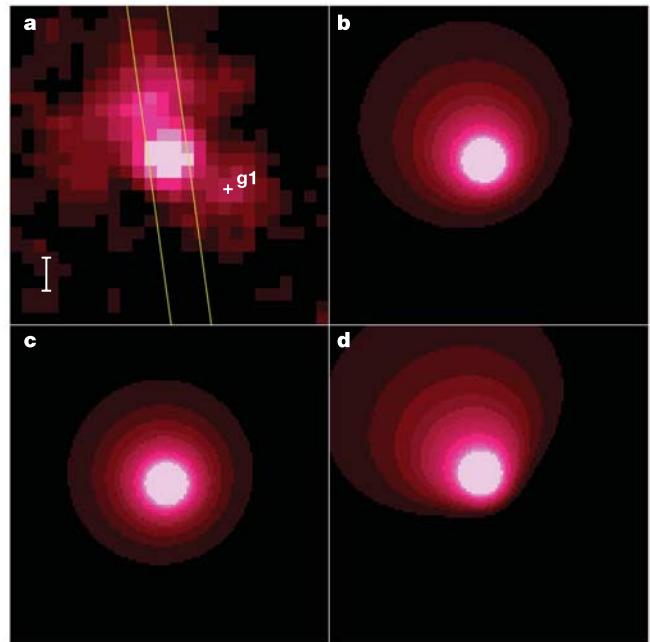


Figure 2 Narrow-band image of the extended Ly α emission compared to models. **a**, Image of the $10 \times 10 \text{ arcsec}^2$ region around Q1205-30 seen in Ly α . The quasar has been subtracted from this image to reveal the underlying extended emission. A 1.2-arcsec -wide slit is overplotted at a position angle of 7.9° east of north, and the position of an unrelated foreground galaxy⁵ is marked g1. The vertical bar on the lower left is 1 arcsec high. **b–d**, Calculated $10 \times 10 \text{ arcsec}^2$ surface brightness maps for an opening angle of $\psi = 110^\circ$ and three inclination angles. The best-fitting inclination angle ($\theta = 27^\circ$) is shown in **b**, two other inclination angles are shown in **c** ($\theta = 10^\circ$, excluded at 3.5σ) and **d** ($\theta = 50^\circ$, excluded at 3σ) for comparison. Given an opening angle, the best-fit inclination angle is determined in the following way. An average surface brightness profile, parallel and perpendicular to the apparent major axis of the extended emission, is fitted to the same average calculated from the models. For each opening angle ψ the best-fitting inclination angle is found.

emission along the line of sight⁹. We take the gas density profile to be a power law, $n_{\text{HI}}(r) = n_{\text{HI},1}(r/1 \text{ kpc})^{-\alpha}$, with a slope α and a neutral hydrogen density at a distance of 1 kpc $n_{\text{HI},1}$, because this appears in numerical simulations of DM halos to be a good approximation at small radii¹⁰. If we assume a mass of the DM halo, we may calculate the observed infall velocity in the model. At a given projected distance from the quasar the observed velocity is the average

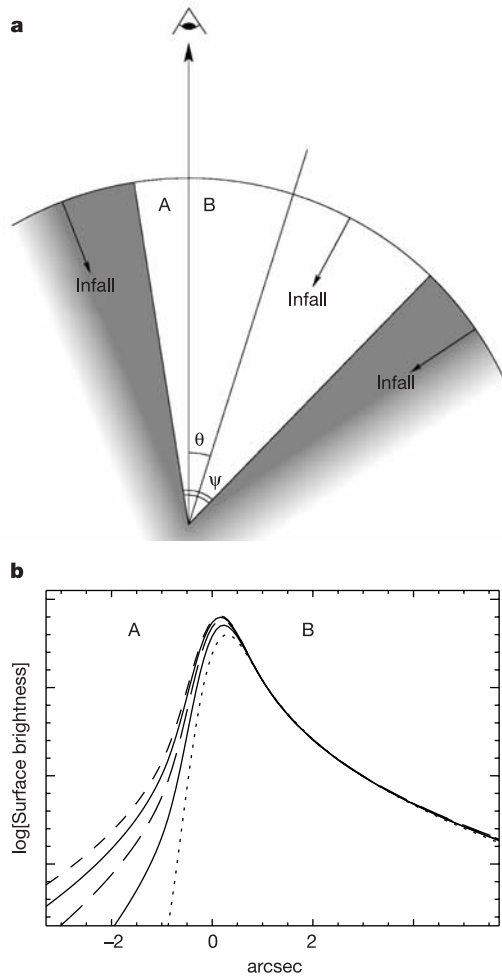


Figure 3 Schematic presentation of the model. **a**, A quasar is located in the centre of a DM halo into which neutral hydrogen is falling. The ionizing photons from the quasar are emitted in a cone of opening angle ψ , and they cause the infalling H I to glow in Ly α . The observer views the system inclined at an angle θ . The surface brightness at each point is calculated by integration of the volume emission along the line of sight⁹:

$$\Sigma_{\text{Ly}\alpha} = \frac{\int E_{\text{Ly}\alpha} \hat{n}_{\text{Ly}\alpha}(r) dl D_A^2}{4\pi D_L^2 d\Omega},$$

where $E_{\text{Ly}\alpha} = 10.4 \text{ eV}$ is the energy of a Ly α photon, $\hat{n}_{\text{Ly}\alpha}(r)$ is the production rate of Ly α photons at a distance r from the quasar, D_L is the quasar luminosity distance, and D_A is the angular distance, such that $d\Omega/D_A^2$ is the conversion from cm^2 to arcsec^2 . The Ly α production rate is $\hat{n}_{\text{Ly}\alpha}(r) = \eta_{\text{thin}} n_{\text{HI}}(r) I(r)$, and $\eta_{\text{thin}} = 0.42$ is the probability for an ionizing photon to result in a Ly α photon in the optically thin case, $n_{\text{HI}}(r)$ is the volume density of neutral hydrogen atoms, and $I(r)$ is the ionization rate of hydrogen atoms. **b**, A schematic illustration of the predicted surface brightness profile for various opening angles $\psi = 90^\circ$ (dotted), 110° (solid), 130° (long-dashed), 150° (solid), and 170° (short-dashed), and a fixed inclination angle $\theta = 40^\circ$. The profiles are presented as they would appear when observed with a seeing of 0.7 arcsec through a slit aligned with the line between the sightline l and the symmetry axis of the cone. Negative angular distances correspond to emission seen from side A, and positive angular distances correspond to emission seen from side B.

projected velocity weighted with the emissivity of the gas along the line of sight.

Applying this model to Q1205-30, our calculation shows that the Ly α halo should be detectable with 8-m class telescopes up to several arcsec from the quasar, assuming a modest amount of neutral hydrogen infall. Considering—for now—the case where only one cone is visible, the extended emission may be symmetric (if $\theta \approx 0$, that is, the cone is seen close to end-on) or highly asymmetric (if $\theta \approx \psi/2$). By fitting the calculated surface brightness maps to the observed narrow-band image we establish a relation between the opening angle and the best-fitting inclination angle. Thus, we are left with the opening angle, ψ , the slope, α , and the neutral hydrogen density at 1 kpc, $n_{\text{HI},1}$, as the only free parameters of the model. In

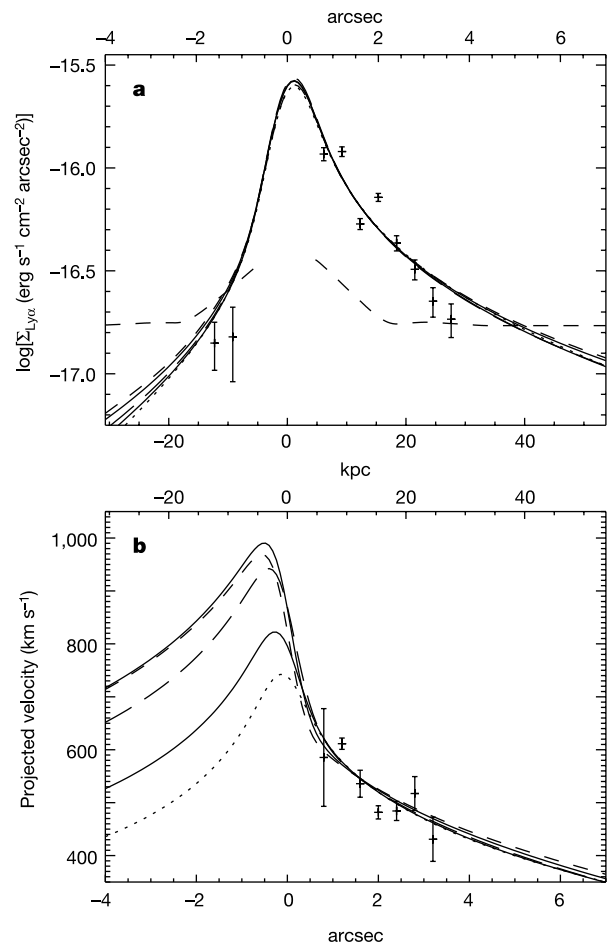


Figure 4 Calculated surface brightness profile and projected velocity. **a**, Comparison of the calculated and observed surface brightness profile. Thin lines are the calculated profile for various opening angles, and data points are our observations of the extended Ly α emission around Q1205-30. The calculations were performed for opening angles and inclination angles of $\psi = 90^\circ$, $\theta = 22.5^\circ$ (dotted), $\psi = 110^\circ$, $\theta = 27.2^\circ$ (solid), $\psi = 130^\circ$, $\theta = 31.8^\circ$ (long-dashed), $\psi = 150^\circ$, $\theta = 36.5^\circ$ (solid), and $\psi = 170^\circ$, $\theta = 41.1^\circ$ (short-dashed). The thick dashed line shows our 5σ detection limit. **b**, Comparison of calculated and observed projected velocity relative to the quasar redshift. Lines are the calculated best-fit velocity profiles for opening angles and inclination angles of $\psi = 90^\circ$, $\theta = 22.5^\circ$ (dotted), $\psi = 110^\circ$, $\theta = 27.2^\circ$ (solid), $\psi = 130^\circ$, $\theta = 31.8^\circ$ (long-dashed), $\psi = 150^\circ$, $\theta = 36.5^\circ$ (solid), and $\psi = 170^\circ$, $\theta = 41.1^\circ$ (short-dashed). Three of the measurements in panel **a** were too faint to allow a secure velocity determination. In both panels the angular distance was converted to physical distance assuming a flat $\Omega_A = 0.7$ Universe with a Hubble constant $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. Error bars are 1σ .

Fig. 2, we compare our observed narrow-band image to calculated Ly α surface brightness maps with an opening angle $\psi = 110^\circ$ and three different inclination angles: the best fit and two others. For the expected large opening angles ($\psi \approx 90^\circ - 120^\circ$)^{11,12}, varying the opening angle primarily affects the shape of the surface brightness profile seen on the side where the emission is weakest (in Fig. 3, that would be on side A of the sightline), and has only a minor effect on the shape of the main emission profile (on side B). The gas density scale, $n_{\text{HI}, 1}$, determines the normalization of the surface brightness profile, but not its shape, which is set by the slope α . The results are not dependent on the choice of a power-law gas density profile. An exponential gas density profile provides an equally good fit to the data.

The extended emission is spatially resolved, so we can measure its surface brightness profile and velocity profile (see Fig. 4). The surface brightness profile is measured by integrating the flux from 4,900 to 4,947 Å in each spatial bin. The velocity in each spatial bin is measured by fitting a gaussian to the line profile. The infall velocity is calculated relative to the redshift $z = 3.041$ of the quasar.

We find that the model describes the observations well. The opening angle is the only parameter which remains unconstrained, so we plot the surface brightness profile for a range of opening angles (Fig. 4a). The best-fit values for the neutral hydrogen density at 1 kpc vary less than a factor of two between $\psi = 90^\circ$ and 170° , while the slope varies between 0.02 and 0.09. The velocity profile of the extended emission (Fig. 4b) arises as a projection effect, and fitting velocity curves of canonical DM halo profiles³ to the observed velocities, we infer a virial mass of $(2 - 7) \times 10^{12} M_\odot$ for the DM halo. An identical mass estimate is obtained when using an exponential gas density profile. The H I density from our best-fitting model allows us to calculate an accretion rate of $\sim 0.1 M_\odot \text{ yr}^{-1}$ in neutral hydrogen. The total accretion rate will be higher as the gas is highly ionized.

Where studies based only on spectra are unable to provide evidence to distinguish between several different scenarios, the combination of deep imaging and deep spectroscopy allows us to rule out alternative explanations of the extended emission. Jets are generally believed to be present in radio-quiet quasars and are predicted to extend out to ~ 0.1 kpc (ref. 13), where the extended Ly α emission around Q1205-30 extends out to ~ 30 kpc. Outflowing galactic winds are believed to be triggered by supernovae going off inside the galaxy, and are therefore expected to be metal-enriched. Our detection limit of $4 \times 10^{-18} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ arcsec}^{-2}$ (3 σ) ensures that we were able to detect the C IV or He II lines typical in extended emission around radio-loud quasars¹⁴. We did not detect any of these lines, so the gas appears not to have been enriched by supernovae. □

Received 24 March; accepted 28 June 2004; doi:10.1038/nature02793.

1. Carlberg, R. G. Quasar evolution via galaxy mergers. *Astrophys. J.* **350**, 505–511 (1990).
2. Haiman, Z. & Rees, M. Extended Ly α emission around young quasars: a constraint on galaxy formation. *Astrophys. J.* **556**, 87–92 (2001).
3. Navarro, J. F., Frenk, C. S. & White, S. D. M. A universal density profile from hierarchical clustering. *Astrophys. J.* **490**, 493–508 (1997).
4. Barkana, R. & Loeb, A. Spectral signature of cosmological infall of gas around the first quasars. *Nature* **421**, 341–343 (2003).
5. Fynbo, J. P. U., Thomsen, B. & Møller, P. Ly α emission from a Lyman limit absorber at $z = 3.036$. *Astron. Astrophys.* **353**, 457–464 (2000).
6. Møller, P. Spectral PSF subtraction I: The SPSF look-up-table method. *ESO Messenger* **99**, 31–33 (2000).
7. Møller, P., Warren, S. J., Fall, S. M., Jakobsen, P. & Fynbo, J. U. SPSF subtraction II: The extended Ly α emission of a radio quiet QSO. *ESO Messenger* **99**, 33–35 (2000).
8. Bunker, A., Smith, J., Spinrad, H., Stern, D. & Warren, S. J. Illuminating protogalaxies? The discovery of extended Lyman- α emission around a QSO at $z = 4.5$. *Astrophys. Space Sci.* **284**, 357–360 (2003).
9. Gould, A. & Weinberg, D. H. Imaging the forest of Lyman limit systems. *Astrophys. J.* **468**, 462–468 (1996).
10. Barkana, R. A model for infall around virialized haloes. *Mon. Not. R. Astron. Soc.* **347**, 59–66 (2004).
11. Lawrence, A. The relative frequency of broad-lined and narrow-lined active galactic nuclei – Implications for unified schemes. *Mon. Not. R. Astron. Soc.* **252**, 586–592 (1991).
12. Elvis, M. A structure for quasars. *Astrophys. J.* **545**, 63–76 (2000).
13. Blundell, K. M., Beasley, A. J. & Bicknell, G. V. A relativistic jet in the radio-quiet quasar PG 1407 + 263. *Astrophys. J.* **591**, L103–L106 (2003).

14. Heckman, T. M., Lehnert, M. D., Miley, G. K. & van Breugel, W. Spectroscopy of spatially extended material around high-redshift radio-loud quasars. *Astrophys. J.* **381**, 373–385 (1991).

Acknowledgements This work was based on observations made with ESO telescopes at the Paranal Observatory.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.W. (mweiding@eso.org).

A ‘checkerboard’ electronic crystal state in lightly hole-doped $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$

T. Hanaguri^{1,2}, C. Lupien³, Y. Kohsaka⁴, D.-H. Lee^{5,6}, M. Azuma^{2,7}, M. Takano⁷, H. Takagi^{1,2,4} & J. C. Davis³

¹Magnetic Materials Laboratory, RIKEN (Institute of Physical and Chemical Research), Wako 351-0198, Japan

²Japan Science and Technology Agency, Kawaguchi 332-0012, Japan

³LASSP, Department of Physics, Cornell University, Ithaca, New York 14853 USA

⁴Department of Advanced Materials Science, University of Tokyo, Kashiwa 277-8651, Japan

⁵Department of Physics, University of California, and

⁶Material Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

⁷Institute for Chemical Research, Kyoto University, Uji 601-0011, Japan

The phase diagram of hole-doped copper oxides shows four different electronic phases existing at zero temperature. Familiar among these are the Mott insulator, high-transition-temperature superconductor and metallic phases. A fourth phase, of unknown identity, occurs at light doping along the zero-temperature bound of the ‘pseudogap’ regime¹. This regime is rich in peculiar electronic phenomena¹, prompting numerous proposals that it contains some form of hidden electronic order. Here we present low-temperature electronic structure imaging studies of a lightly hole-doped copper oxide: $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$. Tunnelling spectroscopy (at energies $|E| > 100$ meV) reveals electron extraction probabilities greatly exceeding those for injection, as anticipated for a doped Mott insulator. However, for $|E| < 100$ meV, the spectrum exhibits a V-shaped energy gap centred on $E = 0$. States within this gap undergo intense spatial modulations, with the spatial correlations of a four CuO_2 -unit-cell square ‘checkerboard’, independent of energy. Intricate atomic-scale electronic structure variations also exist within the checkerboard. These data are consistent with an unanticipated crystalline electronic state, possibly the hidden electronic order, existing in the zero-temperature pseudogap regime of $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$.

Scanning tunnelling microscopy (STM) has recently emerged as a suitable technique to search for ‘hidden’ electronic order in the copper oxides. $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ (Bi-2212) studies where high-transition-temperature (T_c) superconductivity (HTSC) was suppressed to reveal the pseudogap have been especially fruitful. The prototypical study yielded a pseudogap-like conductance spectrum (V-shaped without coherence peaks) associated with a ‘checkerboard’ approximately four CuO_2 unit cells square of local-density-of-states (LDOS) modulations surrounding vortex cores². Similar phenomena were discovered throughout the sample above T_c (ref. 3), and within strongly underdoped nano-regions exhibiting pseudogap-like spectra⁴ (see Fig. 1c). Underdoped

Bi-2212 therefore shows a tendency towards checkerboard electronic modulations when HTSC is suppressed. However, it is unclear whether these checkerboard modulations in Bi-2212 represent a true electronic phase, because they exhibit^{2–4} (1) a variety of doping-dependent incommensurate wavevectors, (2) very weak intensities, and (3) short (~ 8 nm) correlation lengths within the nanoscale disorder^{4–7}. Furthermore, their atomic-scale spatial and energetic structures are unknown, presumably because of disorder^{4–7} and/or thermal energy smearing $\Delta E \approx 3.5k_B T \approx 30$ meV at $T \approx 100$ K (ref. 3).

To search for electronic order hidden in the pseudogap while avoiding these uncertainties, we decided to study a simpler and less disordered copper oxide at lower doping and temperature. We chose $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$ (Na-CCOC), a material whose parent compound $\text{Ca}_2\text{CuO}_2\text{Cl}_2$ is a canonical Mott insulator⁸. Within its undistorted tetragonal crystal structure (Fig. 1b), all the CuO_2 planes are crystallographically identical. Sodium substitution for Ca destroys the Mott insulator state, producing first a nodal metal⁹ in the zero-temperature pseudogap (ZTPG) regime, and eventually HTSC for $x \geq 0.10$ (refs 10, 11). Crucially, Na-CCOC is easily cleavable between CaCl layers to reveal an excellent surface. Initial STM studies showed clean, flat CaCl surfaces (with traces of nanoscale electronic self-organization) which exhibit a V-shaped spectral gap for $|E| < \sim 100$ meV (ref. 12).

Our studies used $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$ samples with Na concentrations $x = 0.08, 0.10$, and 0.12 and bulk $T_c \approx 0, 15$ and 20 K respectively. Atomically flat surfaces are obtained by cleaving below 20 K in the cryogenic ultrahigh vacuum of a dilution refrigerator.

Figure 1d shows a typical topographic image of the CaCl plane with inset showing the quality of atomic resolution achieved throughout. These surfaces exhibit a perfect square lattice, without discernible crystal distortion or surface reconstruction, and with lattice constant a_0 in agreement with X-ray diffraction (3.85 Å).

To image the electronic states in Na-CCOC, we use spatial- and energy-resolved differential tunnelling conductance, $g(\mathbf{r}, E = eV_s)$, measurements from STM. For a strongly correlated system such as a lightly doped Mott insulator, $g(\mathbf{r}, E)$ is proportional to the momentum-space integrated spectral function at $\mathbf{r}$ (ref. 13), rather than $\text{LDOS}(\mathbf{r}, E)$. Nevertheless, $g(\mathbf{r}, E)$ measurements remain a powerful tool for determining atomic-scale spatial rearrangements of electronic structure.

The properties of $g(\mathbf{r}, E)$ should be determined primarily by states in the CuO_2 plane because the CaCl layers are strongly insulating. In support of this, we find that missing surface atoms (arrow in Fig. 1d) do not affect $g(\mathbf{r}, E)$. A typical spatially averaged spectrum $\langle g(\mathbf{r}, E) \rangle$ for $x = 0.12$ is shown in black in Fig. 1c. The high energy conductance for electron extraction is ~ 5 times greater than that for injection and this ratio grows rapidly with falling doping (Supplementary Fig. 1). Such strong bias asymmetries in conductance have long been anticipated. This is because, in a lightly hole-doped Mott insulator, the reservoir of states from which electrons can be extracted at negative sample bias is determined by $1 - p$, while that of hole-states into which electrons can be injected at positive sample bias is determined by p , where p is the number of holes per CuO_2 .

For $|E| < 100$ meV, a slightly skewed V-shaped gap, reaching very

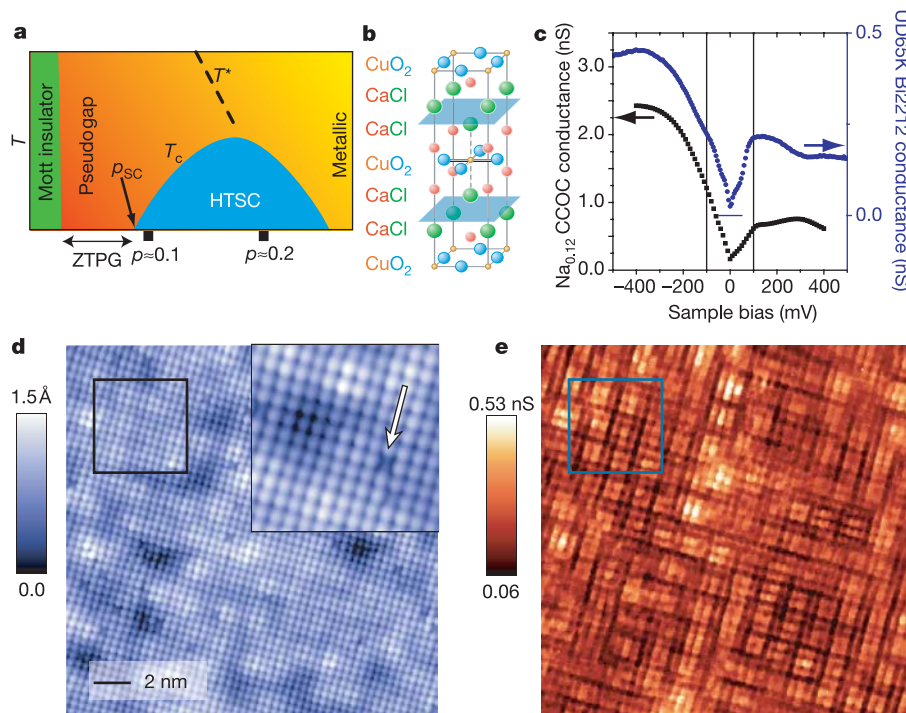


Figure 1 Atomic-scale explorations of electronic states in $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$. **a**, A schematic phase diagram of hole-doped copper-oxides showing the Mott insulator, high- T_c superconductor (HTSC) and metallic phases along with the 'pseudogap' regime and the ZTPG line. **b**, Crystal structure of $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$. Red, orange, blue and green spheres represent Ca (Na), Cu, O and Cl atoms, respectively. Conducting CuO_2 planes are sandwiched by insulating CaCl layers. **c**, A characteristic spatially averaged tunnelling conductance spectrum of $x = 0.12$ Na-CCOC. The large particle-hole asymmetry in conductance at high energies can be associated with the light doping of a Mott insulator (see text). At low energies a skewed V-shaped gap exists. At energies below ~ 10 meV, changes occur in the spectra of superconducting samples (see Supplementary Fig. 5).

The spectrum measured on equivalently underdoped Bi-2212 is shown in blue. **d**, High-resolution STM topograph of the cleaved CaCl plane of a crystal with $x = 0.10$. The perfect square lattice, without discernible bulk or surface crystal reconstructions, is seen. The image was taken at a junction resistance of $4\text{ G}\Omega$ and sample bias voltage $V_s = +200$ mV. **e**, The conductance map $g(\mathbf{r}, E)$ at $E = +24$ meV in the field of view of **d**. It reveals strong modulations with a $4a_0/3 \times 4a_0/3$ commensurate periodicity plus equally intense modulations at $4a_0/3 \times a_0/3$ and strong modulations at $a_0 \times a_0$. All data in this paper are acquired near $T = 100$ mK in a dilution refrigerator-based scanning tunnelling microscope.

low conductance at $E = 0$, is seen in all Na-CCOC spectra (vertical lines in Fig. 1c). Although whether this gap represents precisely the same phenomena as the Bi-2212 pseudogap above T_c (ref. 1) is unknown, it has (modulo thermal smearing effects) very similar basic characteristics: a V-shaped energy gap spanning approximately ± 100 meV with minimum at $E = 0$. Furthermore, tunneling spectra from the most underdoped nano-regions of $p = 0.11 \pm 0.01$ Bi-2212 (blue symbols in Fig. 1c) are almost identical⁴. It seems reasonable that very similar spectra at similar doping in very different lightly doped copper oxides represent the same electronic state. For these reasons, we provisionally designate this spectrum as characteristic of whatever electronic state exists in the ZTPG regime.

We next focus on imaging states within this pseudogap. In Fig. 1e, we show $g(\mathbf{r}, E = 24 \text{ meV})$ measured in the field of view of Fig. 1d. It is reasonably typical of all $g(\mathbf{r}, E)$ obtained for $0.08 \leq x \leq 0.12$ for $10 \text{ meV} < |E| < 100 \text{ meV}$. The most striking fact, and a central result of this paper, is that a clear checkerboard pattern of intense conductance modulations, with primary periodicity $4a_0$, but also complex internal structure at the atomic-scale, is observed.

In Fig. 2a–c we show $g(\mathbf{r}, E)$ measured in the same field of view as Fig. 1d at energies 8, 24 and 48 meV. These and all other subgap $g(\mathbf{r}, E)$ are quite similar. The periodicities of modulations in $g(\mathbf{r}, E)$ are examined via their Fourier transforms $g(\mathbf{q}, E)$, as shown in Fig. 2d–f. These $g(\mathbf{q}, E)$ do not break symmetry under 90° rotations, an observation also noted at $|E| < 100 \text{ meV}$ for all dopings. They

also reveal three primary sets of $\mathbf{q}$ -vectors contributing to the subgap $g(\mathbf{q}, E)$: $\mathbf{q} = (\pm 1, 0)$; $(0, \pm 1)2\pi/a_0$ from the lattice, $\mathbf{q} = (\pm 1/4, 0)$; $(0, \pm 1/4)2\pi/a_0$ from the commensurate $4a_0 \times 4a_0$ modulation, and an unanticipated $\mathbf{q} = (\pm 3/4, 0)$; $(0, \pm 3/4)2\pi/a_0$ modulation, which is also intense. Plots of Fourier intensity along the two orthogonal lines $(2\pi, 0)$ and $(0, 2\pi)$ in Fig. 2g–i show these effects directly. Significantly, these sets of $\mathbf{q}$ -vectors do not disperse with increasing energy and can also be detected as modulations in topographic images (Supplementary Fig. 2). All these observations are consistent with a crystalline electronic state.

To investigate spatial characteristics, we analyse autocorrelation images $A(\mathbf{R}, E) = \langle g(\mathbf{r}, E) g(\mathbf{r} + \mathbf{R}, E) \rangle$. For example, $A(\mathbf{R}, E = 24 \text{ meV})$ is shown in Fig. 3a. It clearly exhibits a basic $4a_0 \times 4a_0$ spatial modulation with a correlation length of $\sim 10a_0$. For brevity, we refer to the $4a_0 \times 4a_0$ unit cell as a ‘tile’. Similar $A(\mathbf{R}, E)$ are found for all $|E| < 100 \text{ meV}$. The $A(\mathbf{R}, E)$ analysis also reveals that the internal electronic structure of representative $4a_0 \times 4a_0$ ‘tiles’ consists of 3×3 intense incommensurate conductance maxima, whose spectral weight is related to the strong $\mathbf{q} = (\pm 3/4, 0)$; $(0, \pm 3/4)2\pi/a_0$ peaks in $g(\mathbf{q}, E)$.

The electronic structure within a representative $4a_0 \times 4a_0$ ‘tile’ is examined directly in Fig. 3b, which shows $g(\mathbf{r}, E = 30 \text{ meV})$ measured within the blue box in Fig. 1e. Comparison between Fig. 3b and the simultaneously acquired topography Fig. 3c shows that the lowest conductance coincides with the perimeter atoms of a $4a_0 \times 4a_0$ ‘tile’, but that the local conductance maxima are not

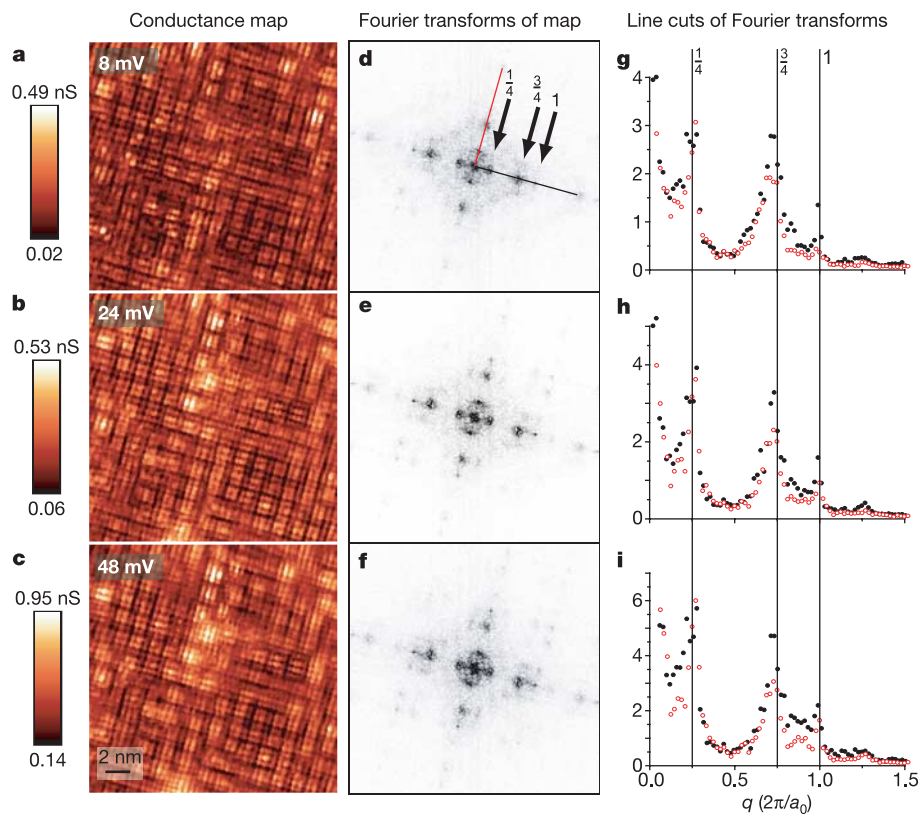


Figure 2 Energy dependence of the modulation wavevectors in the electronic crystal state. **a–c**, Images of differential conductance $g(\mathbf{r}, E)$ at $E = +8, +24$ and $+48 \text{ meV}$ in the same field of view as Fig. 1d; junction resistance was set to $2 \text{ G}\Omega$ at $V_s = +200 \text{ mV}$. Such ‘checkerboard’ patterns are observed at all energies $|E| < 100 \text{ meV}$ with strong correlations between the patterns at each energy. The dark low-conductance one-dimensional streaks are always registered to the atomic rows. The modulations are intense, using up to 90% peak-to-peak of the mean conductance at each energy. **d–f**, The calculated $g(\mathbf{q}, E)$ are Fourier transforms of the $g(\mathbf{r}, E)$ in **a–c**. Regions near the centre of each $g(\mathbf{q}, E)$ reflect the nanoscale electronic disorder. The modulations in

$g(\mathbf{q}, E)$ contain three different $\mathbf{q}$ -vectors: $\mathbf{q} = (\pm 1, 0)$; $(0, \pm 1)2\pi/a_0$, $\mathbf{q} = (\pm 1/4, 0)$; $(0, \pm 1/4)2\pi/a_0$, and $\mathbf{q} = (\pm 3/4, 0)$; $(0, \pm 3/4)2\pi/a_0$. Fourier transforms of these $g(\mathbf{r}, E)$ functions do not break 90° rotational symmetry, within the systematic uncertainty of the measurement as determined by the degree to which the atomic modulations $\mathbf{q} = (\pm 1, 0)$; $(0, \pm 1)2\pi/a_0$ break this symmetry. All three dominant $\mathbf{q}$ -vectors do not disperse with increasing energy. **g–i**, The magnitude of the Fourier transform is plotted along the $(2\pi, 0)$ (black filled circles) and $(0, 2\pi)$ (red open circles) directions for each of these same $g(\mathbf{r}, E)$ and $g(\mathbf{q}, E)$ as in **a–c** and **d–f**.

associated with atomic locations inside the tile (except for the central atom). Figure 3d shows the topographic signal (black) and simultaneously measured conductance (red) along a line through the tile centre.

The energetic changes occurring within a representative 'tile' are even more complex. In Fig. 3e, we show averages of spectra selected from representative regions of low conductance in Fig. 3b (white dots in Fig. 3c) and high conductance (red dots in Fig. 3c—see also Supplementary Fig. 3). From this we see that the pseudogap magnitude varies on the atomic scale by a factor of up to three.

High pseudogap regions have low subgap conductance but also support a strong resonance at $\Omega = 150$ meV (arrow in Fig. 3e).

Next we consider the doping dependence of these phenomena. Bulk electronic transport studies at low temperatures show $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$ to be insulating at $x = 0.06$, marginally conducting at $x = 0.08$, and superconducting for $x \geq 0.10$ (Supplementary Fig. 4). The doping dependence of our subgap $g(\mathbf{r}, E)$ is summarized by Fig. 4. The diminishing bias-voltage asymmetry in conductance spectra with increasing x (Supplementary Fig. 1) indicates that top CuO_2 plane doping increases as expected. On

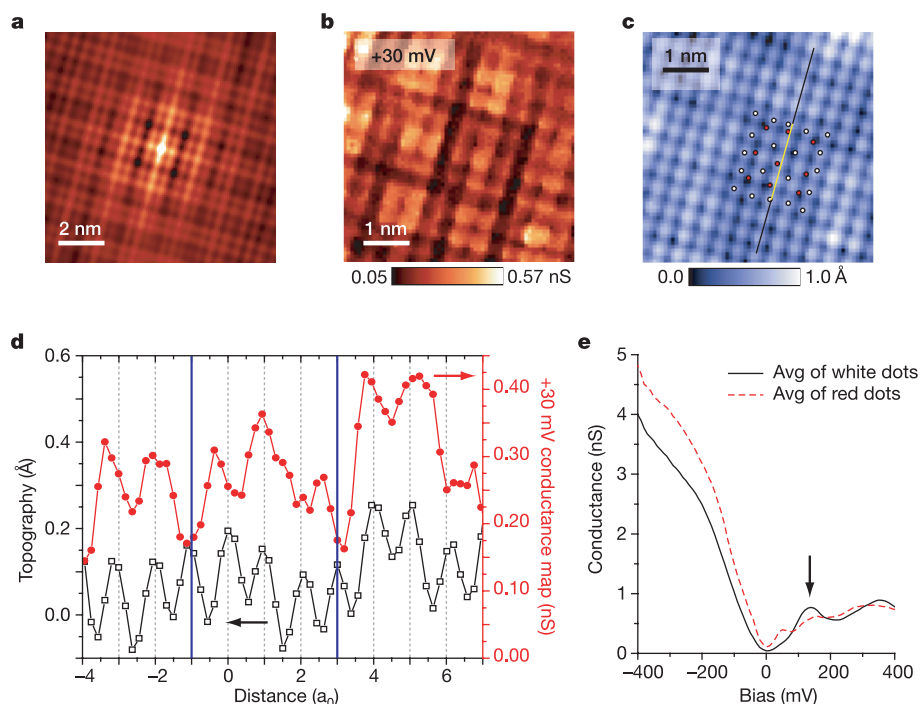


Figure 3 Electronic structure imaging within a representative $4a_0 \times 4a_0$ 'tile'.

a, Autocorrelation image of typical $|E| < 100$ meV LDOS maps showing the $4a_0 \times 4a_0$ structure inside a dark perimeter, which is located exactly where the 16 atoms are positioned at the common perimeter between adjacent $4a_0 \times 4a_0$ regions. **b**, A representative tile of the $4a_0 \times 4a_0$ state is seen directly. The tile exhibits a very low conductance at the perimeter and high LDOS conductance pattern with nine incommensurate maxima inside. The average spatial electronic structure in **a** is remarkably consistent with electronic structure of this (and other) tiles, but this should not be overemphasized because there is also a great deal of variability. Junction resistance for measurement was $2 \text{ G}\Omega$ at $V_s = +400$ mV. **c**, The simultaneously acquired topographic

image showing the locations of what are believed to be the Cl atoms (light) above each Cu atom in the CuO_2 plane. **d**, The topographic signal (black) and simultaneously measured conductance (red) along the line shown in **c** through the tile centre. **e**, Spectra show V-shaped pseudogaps regardless of the positions. The most remarkable difference between dark (low conductance) and bright (high conductance) spots in **b** appears in the positive bias voltage. At dark spots, a strong peak is observed at $\Omega \approx +150$ meV. On average the high pseudogap (most insulating) regions exhibit a strong resonance at $\Omega \approx +150$ meV so that the $g(\mathbf{r}, E = 150 \text{ meV})$ is anticorrelated with the $g(\mathbf{r}, E)$ below 100 meV.

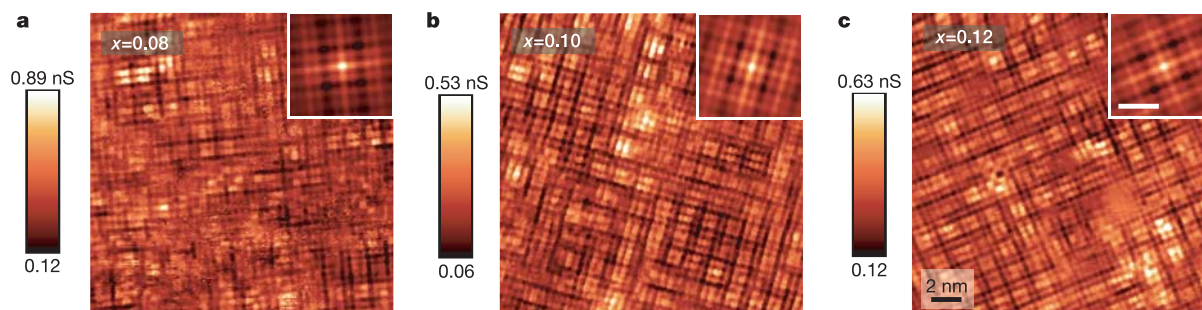


Figure 4 Doping dependence of electronic structure images. **a–c**, The $g(\mathbf{r}, E)$ patterns seen at three different dopings ($x = 0.08, 0.10, 0.12$) are highly characteristic of lightly doped Na-CCOC. Although the high-energy spectra are changing (Supplementary Fig. 1), the subgap spatial characteristics are almost the same in the non-superconducting and

superconducting phases in the energy range $10 \text{ meV} < |E| < 100 \text{ meV}$. The autocorrelation functions seen as insets appear very similar except for a slight increase in correlation length with increasing doping. The white scale bar in the inset corresponds to 2 nm.

the other hand, the basic $4a_0 \times 4a_0$ checkerboard state with its $4/3a_0 \times 4/3a_0$ modulations and pseudogap variations changes subtly so that the $A(\mathbf{R}, E)$ (insets in Fig. 4) appear very similar between dopings. Thus, the checkerboard electronic state appears for $x \leq 0.08$ in the ZTPG regime but continues to exist for $p > p_{SC}$, the critical doping for superconductivity.

We recently introduced a conjecture on underdoped Bi-2212 electronic structure^{4,14} which can help to clarify these observations. In $\mathbf{k}$ -space we distinguish two regions. The first consists of low-energy states on the Fermi arc surrounding the gap nodes. They support a nodal metal at $p < p_{SC}$, which becomes a superconductor at $p > p_{SC}$. Second, the high-energy states near the first Brillouin zone-face are incoherent at low doping owing to localization in an electronic crystal state⁴. At low dopings, these high-energy electronic crystal states are unperturbed by the conversion of low-energy states from a nodal metal to a superconductor. This conjecture predicts that the Na-CCOC checkerboard state would be unperturbed by the transition to superconductivity for dopings near p_{SC} , consistent with observations.

Angle-resolved photoemission spectroscopy (ARPES) studies of Na-CCOC do indeed detect two very different $\mathbf{k}$ -space regions, a large (~ 200 meV) pseudogap around $(\pi, 0)$, plus a coexisting nodal metal or superconductor consisting of states with $|E| < 10$ meV⁹ near the nodes—all consistent with our conjecture. In Bi-2212 the near-nodal states at $|E| < 35$ meV are identifiable with superconductivity via quasiparticle interference⁴. But in $x = 0.12$ Na-CCOC, the near-nodal states occur for $|E| < 10$ meV (ref. 9 and Supplementary Fig. 5), a voltage range where we do not yet have enough sensitivity to detect interference patterns (in any material⁴).

The relationship between the checkerboard electronic crystal state in Na-CCOC and the novel electronic phases proposed for copper oxides^{14–30} is indeterminate. In the absence of disorder, the orbital-current ordered states^{15–17} seem inconsistent with the observations. Although elementary one-dimensional stripes^{18–22} also appear inconsistent because of the 90° rotational symmetry of $g(\mathbf{q}, E)$, defect-dominated stripe states²³ or nematic liquid crystal stripes²⁴ may still be consistent. Checkerboard states due to Fermi surface nesting¹⁴, valence-bond solids^{22,25}, pair density waves^{26,27}, hole-pair^{22,27,28} or single hole^{14,22} Wigner crystals, or the more recently discussed electronic 'supersolids'^{29,30}, are not inconsistent with data herein. However, none of these proposals captures the intricate spatial and electronic characteristics found in $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$. Detailed theory of atomic-scale spectroscopic signatures for the various phases will be required to discriminate between them.

Independently of its microscopic identity, the Na-CCOC checkerboard state is significant because it exemplifies a (no longer hidden) electronic order associated with the copper-oxide pseudogap, and this order is not a translationally invariant liquid of electronic excitations but rather some form of electronic crystal. □

Received 24 February; accepted 13 July 2004; doi:10.1038/nature02861.

- Timusk, T. & Statt, B. The pseudogap in high-temperature superconductors: an experimental survey. *Rep. Prog. Phys.* **62**, 61–122 (1999).
- Hoffman, J. E. *et al.* A four unit cell periodic pattern of quasi-particle states surrounding vortex cores in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Science* **295**, 466–469 (2002).
- Vershinin, M. *et al.* Local ordering in the pseudogap state of the high- T_c superconductor $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Science* **303**, 1995–1998 (2004).
- McElroy, K. *et al.* Destruction of antinodal state coherence via 'checkerboard' charge ordering in strongly underdoped superconducting $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. Preprint at (<http://arXiv.org/cond-mat/0406491>) (2004).
- Howald, C., Fournier, P. & Kapitulnik, A. Inherent inhomogeneities in tunneling spectra of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8-x}$ crystals in the superconducting state. *Phys. Rev. B* **64**, 100504 (2001).
- Cren, T., Roditchev, D., Sacks, W. & Klein, J. Nanometer scale mapping of the density of states in an inhomogeneous superconductor. *Europhys. Lett.* **54**, 84–90 (2001).
- McElroy, K. *et al.* Relating atomic-scale electronic phenomena to wave-like quasiparticle states in superconducting $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Nature* **422**, 592–596 (2003).
- Ronning, F. *et al.* Photoemission evidence for a remnant Fermi surface and a d -wave-like dispersion in insulating $\text{Ca}_2\text{CuO}_2\text{Cl}_2$. *Science* **282**, 2067–2072 (1998).
- Ronning, F. *et al.* Evolution of a metal to insulator transition in $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$ as seen by angle-resolved photoemission. *Phys. Rev. B* **67**, 165101 (2003).

- Hiroi, Z., Kobayashi, N. & Takano, M. Probable hole-doped superconductivity without apical oxygens in $(\text{Ca}, \text{Na})_2\text{CuO}_2\text{Cl}_2$. *Nature* **371**, 139–141 (1994).
- Kohsaka, Y. *et al.* Growth of Na-doped $\text{Ca}_2\text{CuO}_2\text{Cl}_2$ single crystals under high pressures of several GPa. *J. Am. Chem. Soc.* **124**, 12275–12278 (2002).
- Kohsaka, Y. *et al.* Imaging nano-scale electronic inhomogeneity in lightly doped Mott insulator $\text{Ca}_{2-x}\text{Na}_x\text{CuO}_2\text{Cl}_2$. *Phys. Rev. Lett.* (2004) (in the press); preprint at (<http://arXiv.org/cond-mat/0406089>).
- Wei, J. Y. T. *et al.* Quasiparticle tunneling spectra of the high- T_c mercury cuprates: Implications of the d -wave two-dimensional van Hove scenario. *Phys. Rev. B* **57**, 3650–3662 (1998).
- Fu, H. C., Davis, J. C. & Lee, D.-H. On the charge ordering observed by recent STM experiments. Preprint at (<http://arXiv.org/cond-mat/0403001>) (2004).
- Varma, C. M. Non-Fermi-liquid states and pairing instability of a general model of copper oxide metals. *Phys. Rev. B* **55**, 14554–14580 (1997).
- Chakravarty, S., Laughlin, R. B., Morr, D. K. & Nayak, C. Hidden order in the cuprates. *Phys. Rev. B* **63**, 094503 (2001).
- Lee, P. A. Pseudogap and competing states in underdoped cuprates. Preprint at (<http://arXiv.org/cond-mat/0307508>) (2003).
- Zaanen, J. & Gunnarsson, O. Charged magnetic domain lines and the magnetism of high- T_c oxides. *Phys. Rev. B* **40**, 7391–7394 (1989).
- Machida, K. Magnetism in La_2CuO_4 based compounds. *Physica C* **158**, 192–196 (1989).
- White, S. R. & Scalapino, D. J. Phase separation and stripe formation in the two-dimensional t - J model: A comparison of numerical results. *Phys. Rev. B* **61**, 6320–6326 (2000).
- Emery, V. J., Kivelson, S. A. & Tranquada, J. M. Stripe phases in high-temperature superconductors. *Proc. Natl Acad. Sci. USA* **96**, 8814–8817 (1999).
- Sachdev, S. Quantum criticality: Competing ground states in low dimensions. *Science* **288**, 475–480 (2000).
- Bosch, M., van Saarloos, W. & Zaanen, J. Shifting Bragg peaks of cuprate stripes as possible indications for fractionally charged kinks. *Phys. Rev. B* **63**, 092501 (2001).
- Kivelson, S. A., Fradkin, E. & Emery, V. J. Electronic liquid-crystal phases of a doped Mott insulator. *Nature* **393**, 550–553 (1998).
- Vojta, M. Superconducting charge-ordered states in cuprates. *Phys. Rev. B* **66**, 104505 (2002).
- Podolsky, D., Demler, E., Damle, K. & Halperin, B. I. Translational symmetry breaking in the superconducting state of the cuprates: Analysis of the quasiparticle density of states. *Phys. Rev. B* **67**, 094514 (2003).
- Chen, H.-D., Vafek, O., Yazdani, A. & Zhang, S.-C. Pair density wave in the pseudogap state of high temperature superconductors. Preprint at (<http://arXiv.org/cond-mat/0402323>) (2004).
- Chen, H.-D., Hu, J.-P., Capponi, S., Arrighoni, E. & Zhang, S.-C. Antiferromagnetism and hole pair checkerboard in the vortex state of high T_c superconductors. *Phys. Rev. Lett.* **89**, 137004 (2002).
- Tešanović, Z. Charge modulation, spin response, and dual Hofstadter butterfly in high- T_c cuprates. Preprint at (<http://arXiv.org/cond-mat/0405235>) (2004).
- Anderson, P. W. A suggested 4×4 structure in underdoped cuprate superconductors: a Wigner supersolid. Preprint at (<http://arXiv.org/cond-mat/0406038>) (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge and thank P. Coleman, E. Demler, M. Franz, J. E. Hoffman, P. A. Lee, K. Machida, K. McElroy, D. Pines, S. Sachdev, T. Senthil, T. Timusk, M. Vojta and J. Zaanen for discussions and communications. This work was supported by the ONR, NSF, MEXT, JST and NEDO. C.L. acknowledges support from a NSERC Postdoctoral Fellowship and Y.K. from a JPSJ Research Fellowship for Young Scientists.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.C.D. (jcdavis@ccmr.cornell.edu) or T.H. (hanaguri@riken.jp).

Detachment fronts and the onset of dynamic friction

Shmuel M. Rubinstein, Gil Cohen & Jay Fineberg

The Racah Institute of Physics, The Hebrew University of Jerusalem, Givat Ram, Jerusalem, Israel

The dynamics of friction have been studied for hundreds of years, yet many aspects of these everyday processes are not understood. One such aspect is the onset of frictional motion (slip). First described more than 200 years ago as the transition from static to dynamic friction, the onset of slip is central to fields as diverse as physics^{1–3}, tribology^{4,5}, mechanics of earthquakes^{6–11} and fracture^{12–14}. Here we show that the onset of frictional slip is governed by three different types of coherent crack-like fronts: these are

observed by real-time visualization of the net contact area that forms the interface separating two blocks of like material. Two of these fronts, which propagate at subsonic and intersonic velocities, have been the subject of intensive recent interest^{12–17}. We show that a third type of front, which propagates an order of magnitude more slowly, is the dominant mechanism for the rupture of the interface. No overall motion (sliding) of the blocks occurs until either of the slower two fronts traverses the entire interface.

Consider the relative motion of two blocks of material which are separated by a rough interface. A force, F_N , is applied normal to the interface and a shear force, F_S , is applied in a direction parallel to the interface plane. We define 'slip' as differential motion along parts of the interface, while 'sliding' is defined as motion of the entire interface. More than 200 years ago, Coulomb and Amontons showed that the onset of motion occurs when $F_S = \mu_S F_N$, whereas continuous sliding is sustained for $F_S = \mu_D F_N$. The parameters μ_S and μ_D are, respectively, the static and dynamic coefficients of friction, where $\mu_S \geq \mu_D$. The modern study of dry friction (friction without lubrication) is rooted in the work of Bowden and Tabor⁴, who explained this empirical relation as resulting from an increase of the net area of contact along the interface with increasing F_N . If the two blocks are not atomically smooth, the interface is made up of myriad randomly distributed 'micro-contacts', whose total surface area is smaller than the nominal contact area (that is, the cross-sectional contact area of the blocks) by orders of magnitude. When a normal force is first applied, the enormous initial pressure at its tip deforms each micro-contact⁴. This deformation (which may be

either elastic or plastic) increases its contact area, thereby reducing the resultant pressure. These ideas were subsequently generalized by 'rate and state variable constitutive laws'^{3,10,11,16,18,19}, where an empirical state variable is introduced to describe the time dependence of the frictional interface. These laws have been successful in predicting stability criteria of frictional sliding¹⁶ under changing loading conditions and contact history. Both direct calculations¹⁷ as well as Coulomb and rate and state type frictional laws^{7,9,20} have been used of late to model rapid motion within an interface formed by dissimilar materials. An interesting result of this work, in the context of earthquake dynamics^{6,8,21}, is the prediction of self-healing 'slip' pulses that propagate near the Rayleigh wave speed, V_R .

Our experiments study the dynamics of two brittle acrylic (polymethyl-methacrylate, PMMA) blocks separated by a rough interface, consisting of many discrete randomly distributed micro-contacts. Slip was initiated by increasing F_S at a constant rate after setting a constant value of F_N . The experimental system (Fig. 1) was designed to allow light to pass through the interface only at the actual points of contact (see Methods for details). At all other points no light is transmitted, as the incident light undergoes total internal reflection at the interface. Thus, the light intensity, $I(x,y,t)$, passing through any region along the interface is proportional to the net local contact area $A(x,y,t)$. (Directions x and y are defined in Fig. 1a; t is time.) Imaging of the transmitted intensity at sampling rates of up to 250,000 frames s^{-1} enables us to visualize nearly instantaneous changes of $A(x,y,t)$. The linear dependence of the total intensity, I , on both F_N (Fig. 2a) and the frictional force (Fig. 2b), $F_S = \mu_S F_N$ at the onset of slip, demonstrates the utility of the measurement system, and provides further¹⁰ direct validation of the Bowden and Tabor picture of friction.

We use dynamic measurements of $I(x,y,t)$ to study the block detachment process (that is interface rupture) which gives rise to the onset of relative motion (sliding). These dynamics are governed by well-defined crack-like fronts (Fig. 3), which extend in the y direction and propagate in the x direction along the interface. We define these fronts as 'detachment fronts'. Behind a detachment front the net contact area is reduced, whereas ahead the contact area is unchanged. Three distinct types of detachment fronts, each with its own characteristic signature and velocity range, are observed.

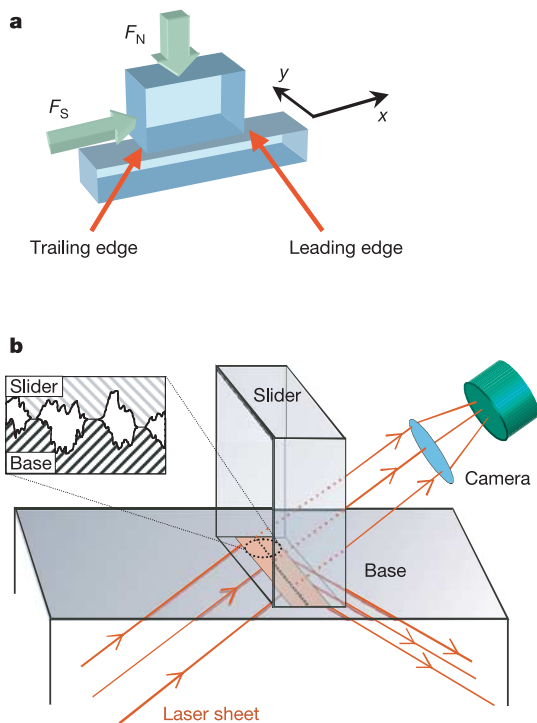


Figure 1 A schematic diagram of the experimental apparatus. **a**, A normal force, F_N , is applied to the base and slider Plexiglas blocks. A shear force, F_S , applied at the trailing edge of the slider, is increased until motion ensues in the x direction. **b**, A laser sheet, incident at an angle beyond the critical angle for total internal reflection, is solely transmitted at the net contact points along the rough interface (inset) between the base and slider. Thus, the light intensity transmitted across the interface at any location is proportional to the net contact area at that location. The transmitted light is imaged by a fast camera (see Methods for details).

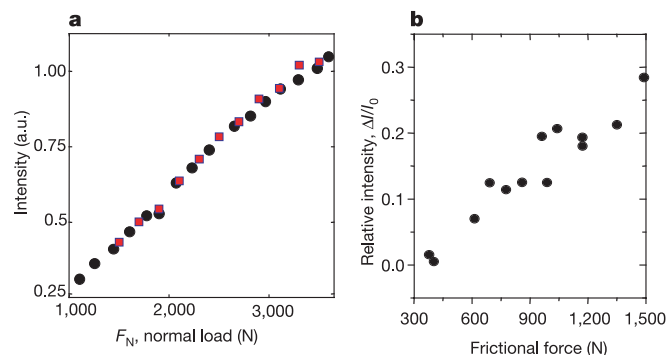


Figure 2 The net contact area of the interface between two blocks of material is proportional to the applied force. **a**, Static measurements of the total intensity of light, I , transmitted across an interface as a function of the applied normal force, F_N . The linear dependence provides experimental verification of the proportionality of net contact area (A) with F_N , as predicted in ref. 4. The different symbols are taken from two different experiments. **b**, The critical shear force at the onset of frictional slip is proportional to transmitted intensity and hence to the net area A across the interface. Shown are measurements from a number of different experiments of the change in transmitted intensity $\Delta I = (I - I_0)$ (immediately before slip), relative to the intensity I_0 of an initial load of $F_N = 750$ N. Each point is the average of four consecutive sliding events, separated in time by a 1 s interval. a.u., arbitrary units.

The detachment process is initiated by a single rapid 'sub-Rayleigh' front¹⁵, which propagates along the interface at velocities that are always less than V_R . This subsonic front accelerates towards V_R and bifurcates into two distinctly different types of detachment fronts; faint 'intersonic' fronts^{12,13} that propagate at velocities that are beyond V_R , and slow detachment fronts that propagate along the interface at velocities over an order of magnitude slower than V_R .

This behaviour is demonstrated in Fig. 4. We measure the relative intensity integrated along the y direction, $I(x,t)/I(x,0)$, which is equal to the relative change of the net local contact area as a function of time. Initially (at $t = 0$) there is no motion along the interface. Interface rupture is always initiated at the trailing edge of the sample, where the shear stress is applied ($x = 0$), owing to a small mechanical torque induced by the loading. A number of distinct propagating fronts are evident in the (x,t) plot shown in Fig. 4a, whose velocity records are presented in Fig. 4b. The initial detachment at $x = 0$ is initiated by the sub-Rayleigh fronts, which then rapidly accelerate until they reach velocities in the vicinity of V_R ($V_R = 930 \text{ m s}^{-1}$ in PMMA). At approximately V_R (see Fig. 4c), this subsonic front arrests and is replaced by both a slow detachment front and intersonic front^{12–14,22}, which are simultaneously emitted. Transitions from sub-Rayleigh to intersonic fronts²³ have been noted, but these studies were not able to observe the accompanying slow detachment fronts. The range of intersonic velocities that we measure are between 1,200 and 1,700 m s^{-1} , with a $\pm 200 \text{ m s}^{-1}$ accuracy. Note that shear cracks with point-like stress field singularities at their tip can propagate at all velocities up to a limiting velocity of V_R . Beyond V_R , only the special velocity of $\sqrt{2}V_S$, where V_S is the shear wave velocity ($V_S = 1,000 \text{ m s}^{-1}$ in PMMA), can support point-like singular modes¹⁵. Velocities between $\sqrt{2}V_S$ and the longitudinal wave speed are allowed if dissipation occurs in a region of finite size¹³.

The velocities (between 40 and 80 m s^{-1}) in Fig. 4b are typical for slow detachment fronts. As Fig. 4c demonstrates, these velocities are over an order of magnitude slower than those of other types of fronts, and 3–6 orders of magnitude faster than the driving velocity. In Fig. 4, the slow front does not reach the sample's leading edge, but undergoes a sharp reverse transition back to sub-Rayleigh fronts. Another example of this reverse transition occurs in Fig. 3 when a

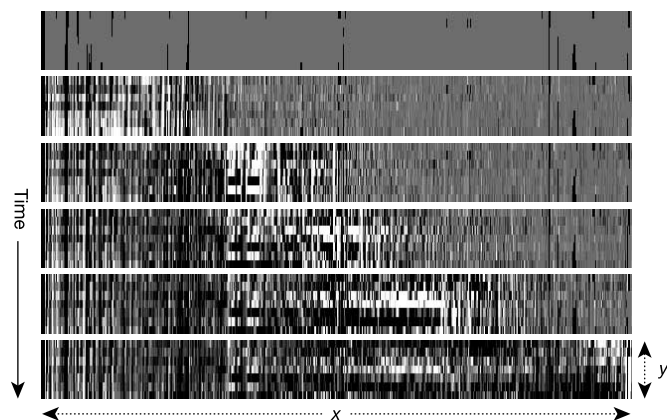


Figure 3 The onset of slip occurs via coherent detachment fronts that propagate across the interface (from left to right). Behind these fronts the net contact area is reduced. Shown are six photographs of the net contact area of the interface between two blocks. The photographs are normalized by the net contact area at $t = 0$. The photographs, having spatial resolution of $1,280 \times 16$ pixels, were taken at (from top to bottom) times of 0, 0.4, 0.75, 1.00, 1.2 and 1.4 ms from the onset of the detachment process. The $(x \times y)$ scales of each photograph are $140 \times 6 \text{ mm}$. Dark (light) shades correspond to a decrease (increase) in contact area.

slow front (frames 2–5) jumps to sub-Rayleigh propagation (between frames 5 and 6). As seen in Fig. 5a, this reverse transition, though common, does not always occur.

An additional feature accompanying the onset of slip is the appearance of 'rebound' waves. Rebound waves, as indicated in Fig. 4a, are initiated at the leading edge of the sample upon the arrival of either sub-Rayleigh or slow detachment fronts. These waves propagate backwards across the interface at intersonic

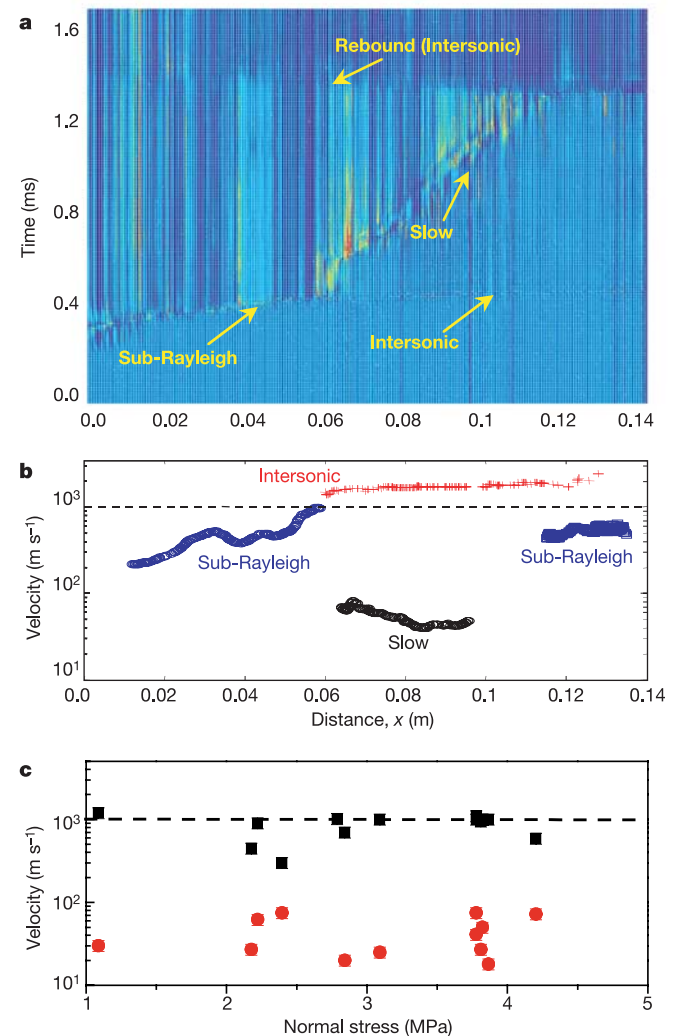


Figure 4 The dynamics of slip, before overall sliding, take place via the interplay between four different types of coherent crack-like fronts. **a**, An (x,t) plot of relative intensity measurements, averaged in the y direction, of a typical experiment for $F_N = 3,470 \text{ N}$, shear loading rate 40 N s^{-1} and driving velocity of $10 \mu\text{m s}^{-1}$ before the onset of slip and a $100,000 \text{ frames s}^{-1}$ sampling rate. The intensity measurements, normalized by their initial values at each point, are colour-coded to reflect the change in the contact area at each spatial point as a function of time. Hot (cold) colours reflect increased (decreased) net contact area. The four different fronts (labelled within the graph) are clearly evident. **b**, Velocity records in space corresponding to the slow, sub-Rayleigh, and intersonic fronts seen in **a**. The rebound front (not shown in **b**) propagated in **a** at a mean velocity of $1,200 \text{ m s}^{-1}$. **c**, The mean velocity values (rectangles) prior and (ovals) subsequent to the transition from sub-Rayleigh to slow fronts as a function of the applied normal stress in 13 different experiments. The velocities shown are mean values over spatial intervals (typically 1 cm) significantly larger than the front position uncertainty. Error bars represent the s.d. of each measurement. Note that the velocities of the two fronts differ by over an order of magnitude, and the mean velocity of the slow detachment fronts is six orders of magnitude more rapid than the rate ($10 \mu\text{m s}^{-1}$) at which the shear stress is applied. In both **b** and **c** the dashed lines indicate V_R .

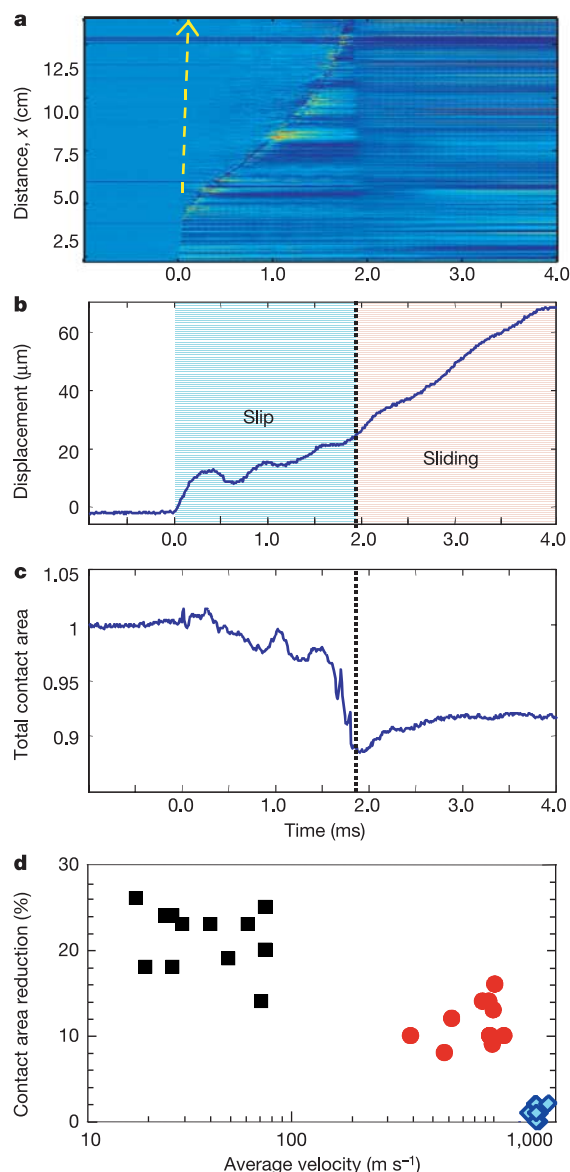


Figure 5 Detailed dynamics of the detachment process. **a**, A (t, x) plot of relative intensity measurements as a function of time for $F_N = 3,510$ N and a shear loading rate of 40 N s^{-1} with a driving velocity of $10 \mu\text{m s}^{-1}$ before the onset of slip. The intensity measurements are averaged in the y direction. Colour coding is as in Fig. 4. The path of the intersonic pulse is highlighted by the dashed arrow. **b**, **c**, Corresponding records of the overall slip of the trailing edge of the sample (the edge where the shear force is applied) (**b**) and the integrated contact area along the entire interface (normalized by the initial contact area) (**c**) as a function of time. Dotted vertical lines in **b** and **c** indicate the arrival time of the slow front at the leading edge of the sample. The passage of the intersonic pulse results in both negligible (smaller than the estimated $10 \mu\text{m}$ intercontact distance) slip and a negligible reduction in contact area along the interface. Only about 30% of the total displacement of the trailing edge of the sample occurs before the arrival of the slow detachment pulse at the leading edge of the sample. At that time, the entire $\sim 12\%$ reduction of the contact area has occurred and only then is slip initiated at the leading edge. **d**, The mean reduction of contact area of the slow detachment fronts (rectangles), sub-Rayleigh fronts (ovals) and intersonic fronts (diamonds) as a function of their mean propagation velocities in 12 different experiments. The contact area reduction of the slow detachment fronts is, on average, twice that of the sub-Rayleigh fronts. The contact area reduction of the intersonic fronts is negligible. All measurements are sampled at 100 kHz . Note that the slight oscillations evident in **b** are due to the compliance of our displacement measurement apparatus (see Methods). This results in a displacement measurement error of $2.5 \mu\text{m}$.

velocities of $1,300 \pm 150 \text{ m s}^{-1}$. While forward-propagating intersonic fronts initiate erratic motion of the contact points along an interface, the process of self-organization of the interface is seen to subside immediately upon the passage of the rebound waves.

Simultaneous measurements of $I(x, t)/I(x, 0)$ (Fig. 5a) and the slip of the trailing edge at $x = 0$ (Fig. 5b) show that significant slip of the trailing edge occurs as the contact area is reduced (Fig. 5c), although the majority of the trailing edge displacement occurs only after the detachment fronts have entirely traversed the sample. No displacement of the leading edge of the sample occurs until immediately following arrival of either slow or sub-Rayleigh fronts (dependent on whether a transition from a slow to a sub-Rayleigh front has occurred). Only then (dotted lines in Fig. 5b, c) does sliding ensue and the system passes from static to 'dynamic' friction. Negligible (smaller than the estimated intercontact distance of $10 \mu\text{m}$) slip results from the passage of the intersonic fronts. The overall 12% reduction in the net contact area at the onset of sliding (Fig. 5c) signals the transition from static to dynamic friction. This is consistent with measured values (10–20%) of $(\mu_s - \mu_D)/\mu_s$ (ref. 3), although μ_D is well-defined only when steady-state sliding persists.

We now consider how efficiently each of the different fronts serves to reduce the net contact area along the interface. In Fig. 5d we present measurements of the relative intensity drop, $\Delta I/I(t = 0)$ (equivalent to a relative change, $\Delta A/A$, of the net contact area), that is precipitated by each type of detachment front. Surprisingly, the intersonic fronts give rise to negligible contact area reduction, with less than 2% of the contact area reduced by their passage. This may explain why these fronts lead to no overall slip. Although intersonic fronts certainly trigger interface fluctuations, they may only separate a small subpopulation of contact points, with the remaining contact points effectively anchoring the interface. Figure 5d further indicates that, in contrast, both the subsonic and slow detachment fronts effectively reduce the local contact area with their passage. The slow fronts are about twice as effective as their sub-Rayleigh counterparts, reducing the contact area by about 20%.

The dynamics of friction have, traditionally, been considered to be governed by processes that occur at slow timescales, when the entire slider is in motion. We have observed here the beginning of the transition to frictional sliding. A full description of these complex processes also involves⁴ timescales much larger than those described here. Detachment fronts, such as those considered in this work, have been too rapid to be detectable in most studies of friction, although such fronts have been recently observed in studies of fault nucleation²⁴ and studies of gelatin sliding on glass²⁵. These rapid processes are fundamental to fault nucleation and the onset of slip in earthquake dynamics. Sub-Rayleigh disturbances, propagating at velocities in the range of $(0.7\text{--}0.8)V_S$ for crustal (shallow) earthquakes and $(0.2\text{--}0.6)V_S$ in deep-focus earthquakes, have been inferred²⁶ from seismic records of a number of large earthquakes. Analysis of seismic data from the recent Izmit earthquake²⁷ suggests that a rupture front may have propagated at intersonic velocities, as observed in recent experiments^{12,13}. Our measurements suggest that slow detachment fronts nearly always occur, either as isolated events or in conjunction with the more rapid modes. The slow fronts give rise to a significant amount of slip while in motion, but their acoustic signatures are significantly weaker than the faster modes. In an earthquake, these waves could therefore be masked, but could become apparent if a slip event were arrested before the transition to a rapid earthquake. Recent observations of slow or 'silent' earthquakes²⁸ in which significant slip was observed with a minimal acoustic signature^{29,30} may, therefore, be analogous to these waves.

It is not yet clear if these fronts are related to self-healing pulses^{6–9,11,17}, which are predicted to propagate along bi-material interfaces, while generating spatially localized interface separation. Although contact area reduction is certainly observed, our measure-

ments show no indication of full interface separation at scales larger than a pixel. □

Methods

The experiment was conducted using two PMMA (Plexiglas) blocks, a slider of size $150 \times 75 \times 6$ mm and a base of size $300 \times 270 \times 30$ mm in the x (propagation), y (sample thickness) and z (normal loading) directions, respectively. The slider contact face was diamond-machined to a flatness of better than $0.1 \mu\text{m}$ (r.m.s.) and then roughened to $0.4\text{--}1 \mu\text{m}$ (r.m.s.). The base was also roughened to $1 \mu\text{m}$ (r.m.s.), thus defining the roughness of the interface to be approximately $1 \mu\text{m}$.

Normal loading, F_N , was applied to the slider via a spring array (with a $4 \times 10^5 \text{ N m}^{-1}$ stiffness), ensuring uniform stress distribution. The shear force, F_S , was applied by coupling a stepping motor to the slider via a load cell having a stiffness of $4 \times 10^6 \text{ N m}^{-1}$. After the onset of slip, F_S decreased by 5–8% as a result of this compliance, which is accompanied by a damped oscillation of the applied shear stress (having a maximum amplitude of about $0.005F_S$). The stepping motor moved in discrete steps of $0.04 \mu\text{m}$ with loading rates ranging from $1 \mu\text{m s}^{-1}$ to 10 mm s^{-1} . Shear was applied to the sample 6.5 mm above the interface, thus introducing a small mechanical torque ($2.6 \times 10^{-3} \text{ F}_N \text{ N m}$) in the loading. This resulted in a 2.5% variation of F_N over the length of the interface. Applied normal and shear stresses ranged from 0 to 6 MPa and 0 to 3 MPa, respectively. Both applied forces and trailing edge displacement were measured by means of S-beam load cells. Measurement resolution of the displacement was $0.5 \mu\text{m}$. This was further limited at the onset of slip by a resonant ($2.5 \mu\text{m}$) vibration of the system. Data acquisition was triggered by means of an acoustic sensor coupled to the leading edge of the slider.

The interface was illuminated by a $200 \times 5 \text{ mm}$ laser sheet through the base block at an angle significantly larger than the critical angle for total internal reflection from a PMMA/air interface. The (660 nm) laser light could transverse the slider–base interface either directly, at the micro-contact points between the two blocks, or by evanescence, where incident light (angle of incidence of 70°) tunnels across the air gap between the two blocks. The large ratio between the roughness ($1 \mu\text{m}$ r.m.s.) of the interface and the exponential decay length ($\sim 50 \text{ nm}$) of the evanescent light enables us (over the range of normal forces applied) to neglect evanescent contributions to the intensity passing through the interface. Thus, the light passing through the interface is proportional to the area of net contact (see Fig. 2), where we consider ‘net contact’ as separation of less than 50 nm . A prism array was used to focus the transmitted light directly onto a fast CMOS sensor (VDS CMC1300 camera), enabling rapid imaging of the transmitted intensity. The sensor can be configured to frame sizes of $1,280 \times N$ pixels with frame rates of $500,000/(N+1) \text{ frames s}^{-1}$. In our experiments, the entire interface was measured; each pixel was mapped using cylindrical optics to either $117 \mu\text{m} \times 1.5 \text{ mm}$ ($\Delta x \times \Delta y$) at $100,000 \text{ frames s}^{-1}$ ($N=4$) or $117 \mu\text{m} \times 370 \mu\text{m}$ ($\Delta x \times \Delta y$) at $30,000 \text{ frames s}^{-1}$ ($N=16$). These pixels encompass numerous contact points (we estimate the mean distance between contact points to be about $10 \mu\text{m}$). All images presented were normalized pixel by pixel by the net contact area at time $t=0$. In addition, to avoid noise-dominated weighting of the normalized data, pixels whose intensity in the reference image was lower than 1% of the maximal pixel intensity were excluded from all subsequent analyses.

Determination of the propagation velocities of the different fronts was performed as follows. The intensity measurements were first averaged in the y direction to create a single array of intensities along the x direction for each frame. Using these data, we obtained a time series of intensities measured at each spatial point (Figs 4a and 5a). To each time series one-dimensional wavelet transforms (Haar and db4 kernels) were applied to determine the arrival time of each of the different fronts at each spatial location along the interface (Fig. 4b). Differentiation of these data then yielded the velocity of each front.

Received 7 April; accepted 22 June 2004; doi:10.1038/nature02830.

- Ciliberto, S. & Laroche, C. Energy dissipation in solid friction. *Eur. Phys. J. B* **9**, 551–558 (1999).
- Muser, M. H., Wenning, L. & Robbins, M. O. Simple microscopic theory of Amontons’s laws for static friction. *Phys. Rev. Lett.* **86**, 1295–1298 (2001).
- Baumberger, T., Berthoud, P. & Caroli, C. Physical analysis of the state- and rate-dependent friction law. II. Dynamic friction. *Phys. Rev. B* **60**, 3928–3939 (1999).
- Bowden, F. P. & Tabor, D. *The Friction and Lubrication of Solids* (Oxford Univ. Press, New York, 2001).
- Persson, B. N. J. *Sliding Friction Physical Principles and Applications* (Springer, New York, 2000).
- Heaton, T. H. Evidence for and implications of self-healing pulses of slip in earthquake rupture. *Phys. Earth Planet. Inter.* **64**, 1–20 (1990).
- Cochard, A. & Rice, J. R. Fault rupture between dissimilar materials: Ill-posedness, regularization, and slip-pulse response. *J. Geophys. Res.* **B 105**, 25891–25907 (2000).
- Ben-Zion, Y. Dynamic ruptures in recent models of earthquake faults. *J. Mech. Phys. Solids* **49**, 2209–2244 (2001).
- Andrews, D. J. & Ben-Zion, Y. Wrinkle-like slip pulse on a fault between different materials. *J. Geophys. Res.* **B 102**, 553–571 (1997).
- Dieterich, J. H. & Kilgore, B. D. Direct observation of frictional contacts — new insights for state-dependent properties. *Pure Appl. Geophys.* **143**, 283–302 (1994).
- Rice, J. R., Lapusta, N. & Ranjith, K. Rate and state dependent friction and the stability of sliding between elastically deformable solids. *J. Mech. Phys. Solids* **49**, 1865–1898 (2001).
- Rosakis, A. J., Samudrala, O. & Coker, D. Cracks faster than the shear wave speed. *Science* **284**, 1337–1340 (1999).
- Xia, K., Rosakis, A. J. & Kanamori, H. Laboratory earthquakes: The sub-Raleigh-to-supershear rupture transition. *Science* **303**, 1859–1861 (2004).
- Gao, H. J., Huang, Y. G. & Abraham, F. F. Continuum and atomistic studies of intersonic crack propagation. *J. Mech. Phys. Solids* **49**, 2113–2132 (2001).

- Freund, L. B. The mechanics of dynamic shear crack propagation. *J. Geophys. Res.* **84**, 2199–2209 (1979).
- Scholz, C. H. Earthquakes and friction laws. *Nature* **391**, 37–42 (1998).
- Gerde, E. & Marder, M. Friction and fracture. *Nature* **413**, 285–288 (2001).
- Rice, J. R. & Ruina, A. L. Stability of steady frictional slipping. *J. Appl. Mech.* **50**, 343–349 (1983).
- Kilgore, B. D., Blanpied, M. L. & Dieterich, J. H. Velocity dependent friction of granite over a wide range of conditions. *Geophys. Res. Lett.* **20**, 903–906 (1993).
- Ben-Zion, Y. & Huang, Y. Q. Dynamic rupture on an interface between a compliant fault zone layer and a stiffer surrounding solid. *J. Geophys. Res.* **B 107**, doi:10.1029/2001JB000254 (2002).
- Anooshehpour, A. & Brune, J. N. Wrinkle-like Weertman pulse at the interface between two blocks of foam rubber with different velocities. *Geophys. Res. Lett.* **26**, 2025–2028 (1999).
- Bodin, P., Brown, S. & Matheson, D. Laboratory observations of fault-normal vibrations during stick slip. *J. Geophys. Res.* **B 103**, 29931–29944 (1998).
- Andrews, D. J. Rupture velocity of plane strain shear cracks. *J. Geophys. Res.* **B 81**, 5679–5687 (1976).
- Ohnaka, M. & Shen, L. F. Scaling of the shear rupture process from nucleation to dynamic propagation: Implications of geometric irregularity of the rupturing surfaces. *J. Geophys. Res.* **B 104**, 817–844 (1999).
- Baumberger, T., Caroli, C. & Ronsin, O. Self-healing slip pulses and the friction of gelatin gels. *Eur. Phys. J. E* **11**, 85–93 (2003).
- Kanamori, H., Anderson, D. L. & Heaton, T. H. Frictional melting during the rupture of the 1994 Bolivian earthquake. *Science* **279**, 839–842 (1998).
- Bouchon, M. *et al.* How fast is rupture during an earthquake? New insights from the 1999 Turkey earthquakes. *Geophys. Res. Lett.* **28**, 2723–2726 (2001).
- Crescentini, L., Amoroso, A. & Scarpa, R. Constraints on slow earthquake dynamics from a swarm in central Italy. *Science* **286**, 2132–2134 (1999).
- Linde, A. T. & Sacks, I. S. Slow earthquakes and great earthquakes along the Nankai trough. *Earth Planet. Sci. Lett.* **203**, 265–275 (2002).
- Rogers, G. & Dragert, H. Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip. *Science* **300**, 1942–1943 (2003).

Acknowledgements We thank Z. Reches, A. Sagy and M. Shay for comments. This research was supported by the Israel Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.F. (jay@vms.huji.ac.il).

Ultrahigh-quality silicon carbide single crystals

Daisuke Nakamura¹, Itaru Gunjishima¹, Satoshi Yamaguchi¹, Tadashi Ito¹, Atsuto Okamoto¹, Hiroyuki Kondo², Shoichi Onda² & Kazumasa Takatori¹

¹Toyota Central R&D Laboratories, Inc., Nagakute, Aichi, 480-1192, Japan

²Research Laboratories, DENSO Corporation, 500-1, Nissin, Aichi, 470-0111, Japan

Silicon carbide (SiC) has a range of useful physical, mechanical and electronic properties that make it a promising material for next-generation electronic devices^{1,2}. Careful consideration of the thermal conditions^{3–6} in which SiC {0001} is grown has resulted in improvements in crystal diameter and quality: the quantity of macroscopic defects such as hollow core dislocations (micropipes)^{7–9}, inclusions, small-angle boundaries and long-range lattice warp has been reduced^{10,11}. But some macroscopic defects (about $1\text{--}10 \text{ cm}^{-2}$) and a large density of elementary dislocations ($\sim 10^4 \text{ cm}^{-2}$), such as edge, basal plane and screw dislocations, remain within the crystal, and have so far prevented the realization of high-efficiency, reliable electronic devices in SiC (refs 12–16). Here we report a method, inspired by the dislocation structure of SiC grown perpendicular to the c -axis (a -face growth)¹⁷, to reduce the number of dislocations in SiC single crystals by two to three orders of magnitude, rendering them virtually dislocation-free. These substrates will promote the development of high-power SiC devices and reduce energy losses of the resulting electrical systems.

Single crystals of the conventional electronic materials silicon and gallium arsenide are grown dislocation-free from molten sources by means of the 'necking' process^{18,19}. But SiC single crystals are usually produced by a gas-phase growth method²⁰ in which the necking process is not suitable, because rapid increase of crystal diameter is impossible.

Once grown, a single crystal of SiC is usually produced by means of *c*-face growth, which involves growing the crystal along the {0001} (*c*-axis) direction using a seed of {0001} substrate within an offset angle about 10°. The generation of defects in a *c*-face growth crystal is considerably affected by defects in the seed crystal. Thus, it is crucial not only to optimize the growth conditions but also to reduce the dislocations in the seed crystal. We have noticed the particular dislocation structure of SiC single crystal obtained by *a*-face growth¹⁷, and this is the key to solving the problem. In this Letter, both {1120} and {1100} faces are called '*a*-face', and both <1120> and <1100> axes are called '*a*-axis', because the dislocation structures of crystal grown along both <1120> and <1100> directions are almost the same (strictly speaking, the *a*-axis is not <1100> but <1120>).

First, we investigated the X-ray incident angle (ω) scan rocking curves of the 0004 diffraction peak spectrum obtained from 4H-SiC

{0001} substrate; the substrate was sliced from an *a*-face {1100} growth ingot obtained using a *c*-face growth seed crystal (see Supplementary Information). From this analysis, we found that *c*-axis perturbation parallel to the growth direction is much smaller than that perpendicular to the growth direction. The width of the peak spectrum (full-width at half-maximum, FWHM, 27 arcsec), which was obtained in the case that the ω -scan axis is perpendicular to the growth direction (case A), almost equalled that of perfect crystal (26 arcsec) in our apparatus, indicating that there is little perturbation in the direction parallel to the growth direction of *a*-face growth crystal. On the other hand, the FWHM of the peak which was obtained where the ω -scan axis is parallel to the growth direction (case B) was much broader¹⁷. We surmise that anisotropy with regard to the crystallographic axis perturbation is peculiar to *a*-face growth of hexagonal crystals. Moreover, we found that the FWHM (in case B) plots for various measurement points *z*, which is the distance from seed crystal in the direction parallel to the growth direction, does not change as a function of *z*. In other words, *a*-face growth crystal faithfully inherits only the *c*-axis variations perpendicular to the growth direction. We call the resultant crystal lattice distortion a 'corrugated plate' lattice.

From these results, we propose that most of the dislocations in *a*-face growth SiC single crystals exist almost parallel to the growth direction with the Burgers vectors perpendicular and parallel to the growth direction, and the density of dislocations hardly changes at all with changes in growth height. That is, the 'corrugated plate' lattice is due to a high density of edge dislocations parallel to the growth direction. X-ray topographic analysis found that many edge dislocations exist parallel to the growth direction, and so has confirmed that this model is appropriate (see Supplementary Information). The same result was obtained with the *g* vector (the diffraction vector) being 0004 as with 1120. These results are consistent with our model of dislocation structure. In addition, almost the same results were obtained from another *a*-face growth crystal—{1120}, perpendicular to {1100}.

We attempted to reduce dislocations by utilizing *a*-face growth and the dislocation structure of an *a*-face growth crystal. The process to eliminate dislocations is as follows. Step 1: *a*-face ({1120} or {1100}) growth along the *a*-axis (<1120> or <1100>)

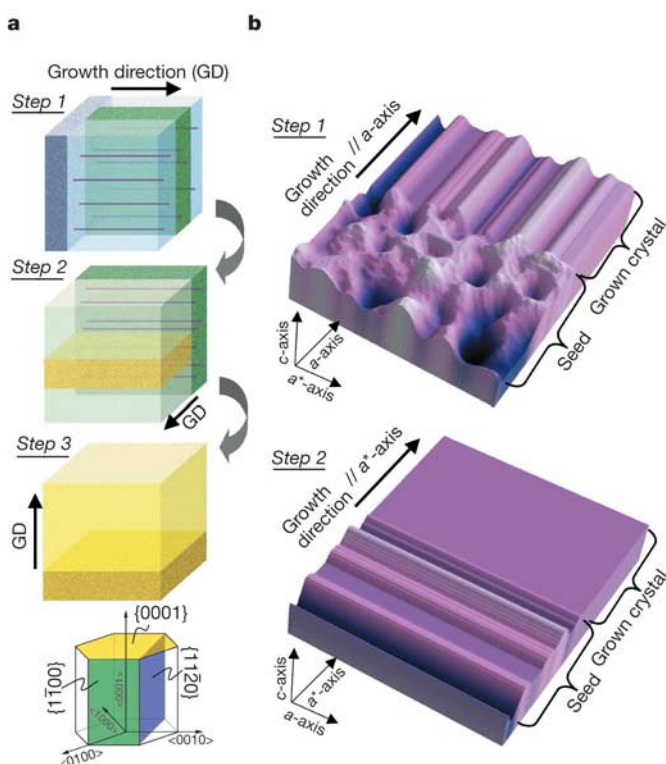


Figure 1 Schematic illustrations of 'repeated *a*-face' (RAF) growth process. The growth sequences are as follows. Step 1: first *a*-face growth (seed and grown crystal are shown dark blue and light blue, respectively). GD, growth direction. Step 2: second *a*-face growth perpendicular to first *a*-face growth (seed sliced from first *a*-face growth crystal, and the grown crystal are shown dark green and light green, respectively). Step 3: *c*-face growth with offset angle of several degrees (seed sliced from second *a*-face growth crystal, and the grown crystal are shown dark yellow and light yellow, respectively). **a**, In this figure, the first and second *a*-faces are {1120} and {1100} faces. At the bottom of panel **a** are shown the major crystallographic axes and lattice planes in the RAF process. Dislocation characteristic of *a*-face growth crystal (first *a*-face growth crystal and second seed crystal of panel **a**) is shown by purple line. **b**, Top side view, showing {0001} lattice plane irregularities of seed and grown crystal in steps 1 and 2. The *a**-axis is perpendicular to both the *a*-axis and the *c*-axis.

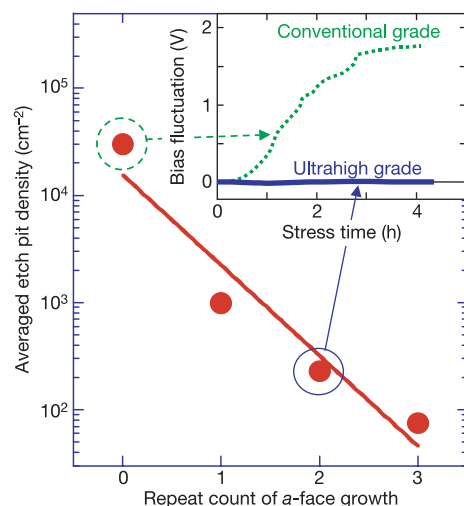


Figure 2 Etch pit density (EPD) decay curves versus repeat count of *a*-face growth, and results of current stress testing. Each averaged EPD was calculated from the results of 90 observation points within a range 20 mm in diameter. The inset shows forward bias fluctuation of PIN diodes under the stress with a constant current density. The broken green line is that of crude substrate²² with a current density of 160 A cm⁻², and the solid blue line is that of RAF substrate with 300 A cm⁻².

direction, using a seed sliced from a *c*-face growth ingot. Step *N* (*N* = 2,3,4,...): *a*-face ($\{1\bar{1}00\}$ or $\{11\bar{2}0\}$) growth along the *a*-axis ($\langle 1\bar{1}00 \rangle$ or $\langle 11\bar{2}0 \rangle$) direction, using a seed sliced from the *a*-face growth ingot of the previous step (step *N* - 1): the seed surface orientation is perpendicular to both the previous step's growth and $\langle 0001 \rangle$ directions. Step *N* + 1: *c*-face growth, using a seed sliced from the *a*-face growth ingot of the previous step (step *N*): the seed surface orientation is $\{0001\}$, with several degrees off-axis towards the perpendicular to both the previous step's growth and $\langle 0001 \rangle$ directions.

We call this the 'repeated *a*-face' (RAF) growth process, because the steps of repeated *a*-face growth are crucial. To simplify, we here explain the case of *N* = 2 (Fig. 1a), where *N* is a repeat count of *a*-face growth. In step 1, we can obtain a crystal with a high density of dislocations, which are inherited from the crude seed crystal, parallel to the growth direction. In step 2, most of the dislocations are not exposed to a surface of the second seed, because most of the dislocations exist perpendicular to the first *a*-face growth direction, that is, parallel to the second seed surface. Therefore, the second *a*-face growth crystal inherits fewer dislocations, and we can obtain a crystal with much lower dislocation density. In addition, it may be seen that we can reduce the dislocation density further by increasing the repeat count *N*. In step 3, we have to eliminate stacking faults¹⁷ due to *a*-face growth in the step 2 growth crystal. We can eliminate such faults by *c*-face growth, because the stacking faults and partial dislocations accompanying them are inherited only perpendicular to the *c*-axis. Moreover, we can reduce the inheritance of dislocations by using a seed with several degrees off-axis towards the perpendicular to both the second *a*-face growth and $\langle 0001 \rangle$ directions, because most of the remaining dislocations in the second *a*-face growth crystal are parallel to the third seed surface. In addition, an offset angle of several degrees is needed to achieve polytype stability²¹.

Figure 1b is a schematic illustration of crystal lattice plane distortion. The $\{0001\}$ lattice plane of *c*-face growth crude seed

crystal is ridged randomly. However, that of the first *a*-face (step 1) growth crystal is like 'corrugated plate' because the *a*-face growth crystal inherits only the *c*-axis perturbation perpendicular to the growth direction from the crude seed crystal surface. Finally, the $\{0001\}$ lattice plane of the second *a*-face (step 2) growth crystal is very flat (in other words, the lattice distortion accompanied by many dislocations is eliminated), because there is little perturbation inherited from the first *a*-face growth seed surface. By experiment, we confirmed that the FWHM values of X-ray rocking curves obtained from second *a*-face growth crystal in both case A and case B were almost equal to each other, both being uniformly low. Moreover, long-range warp of the lattice plane was eliminated with the elimination of *c*-axis perturbation.

Figure 2 shows plots of averaged etch pit density (EPD) of 4H-SiC (0001) 8° off-axis substrates obtained by the RAF process. The repeat count value of zero refers to the *c*-face growth crude crystal. Repeat counts one, two and three refer respectively to the ' $\{1\bar{1}00\}$ and *c*-face', ' $\{1\bar{1}00\}$ and $\{11\bar{2}0\}$ and *c*-face' and ' $\{1\bar{1}00\}$ and $\{11\bar{2}0\}$ and *c*-face' growth crystal in the RAF process. As shown clearly in Fig. 2, the EPDs decrease exponentially with increase in the repeat count of *a*-face growth. This shows that dislocations in the SiC crystal are effectively eliminated by the RAF growth. Moreover, we confirmed by successive etch pit patterns and X-ray topographies that the stacking faults and partial dislocations exposed on the RAF seed surfaces are eliminated perfectly by the *c*-face growth because the stacking faults and partial dislocations propagate only along the $\{0001\}$ plane. The average EPD and micropipe density of a 20-mm-diameter substrate, taken from the crystal grown on RAF seed with *a*-face growth performed three times (Fig. 2), were respectively 75 cm^{-2} and 0 cm^{-2} . This EPD value is lower by three orders of magnitude than that of conventional grade SiC substrates. We think that a repeat count of *a*-face growth greater than three will lead to a lower EPD value and a higher crystal quality; the experiments are now under way. In addition, we have investigated current-induced degradation of PiN devices fabricated on the RAF substrate. As shown in an inset of Fig. 2, only a little fluctuation is observed in forward bias ($<15 \text{ mV}$) on the ultrahigh-quality RAF substrate, whereas a rapid increase occurs in the bias on the conventional grade substrate²². This means that the reliability of bipolar devices is much enhanced by the RAF process.

Next, we attempted to enlarge the crystal diameter by the RAF process. From the standpoint of crystal growth, the enlargement of crystal size give rise to various problems. In the case of SiC crystal growth, the effects of thermal stress and foreign polytype inclusion are most crucial. To maintain good crystal quality under crystal growth, we have optimized the crucible structures, growth conditions and seed positioning method, and have managed both enlargement and quality. As a result, we have obtained large RAF substrates, 1.5–3.0 inches in diameter, whose averaged EPDs were lower by about two orders of magnitude than those of conventional SiC substrates. We have evaluated the crystal quality of the large RAF substrate by synchrotron monochromatic beam X-ray topography (SMBXT) (Fig. 3). As shown in Fig. 3a, crystal quality of the RAF substrate is very homogeneous, and there are very few macroscopic defects and dislocations; we also found that the long-range lattice warp is very small, with the curvature radius of the lattice evaluated to be about 800 m. On the other hand, as shown in Fig. 3b, conventional grade substrate has many macroscopic defects and dislocation networks.

The crystal quality of RAF substrates is much better than that of conventional substrates; we have shown that the RAF growth process is very effective in reducing dislocations and other defects. Moreover, we have succeeded in manufacturing a large size substrate by this method, which makes feasible commercial applications. We consider that it will be possible in the near future to eliminate dislocations perfectly, and to enlarge the diameter to several inches. Finally, we suggest that the concept of the RAF

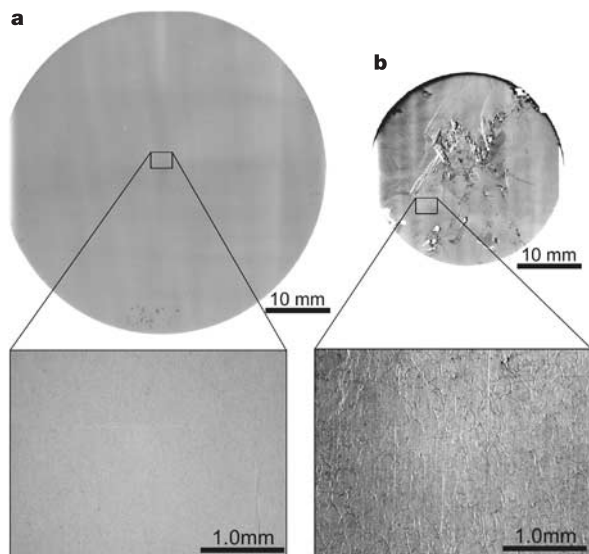


Figure 3 Synchrotron monochromatic beam X-ray topographies. 4H-SiC (0001) 8° off-axis substrate, 2.0 inches in diameter, manufactured by the RAF process (**a**) and a 1.2-inch-diameter specimen manufactured by the conventional process (only *c*-face growth; **b**). Averaged EPD and micropipe density of the RAF growth crystal were about 250 cm^{-2} and 0 cm^{-2} , respectively, and those of crystal grown by the conventional method were about $3 \times 10^4 \text{ cm}^{-2}$ and 30 cm^{-2} , respectively. Enlarged images are shown below. Dislocation networks are shown in the enlarged image of **b**.

process can be applied to single-crystal growth of other materials that have hexagonality in their crystal structures. □

Methods

EPD evaluation by molten KOH etching

EPDs have been considered to be the density of dislocations exposed at the substrate surface. The molten KOH etching was performed at 773 K for 20–30 min in a nickel crucible. After the etching, etch pits were observed by optical microscopy. The field of view at each observation point was $870 \times 650 \mu\text{m}$. The observation points were located in tetragonal grids, and the distance between each point was 2.0 or 3.0 mm.

Synchrotron monochromatic beam X-ray topography

The experiment was done at Spring-8 (BL20-B2). The topographies were taken with Berg-Barrett geometry with the diffraction plane of 1128. Incident X-ray energy was 11.94 keV, and immersion depth of reflection beam was about $4 \mu\text{m}$. The topographies in Fig. 3 were obtained by continuous scan of incident angle (ω), and the zebra patterns to evaluate the curvature radius of crystal lattice planes were taken by step scan of that. Dislocations in the topography appear as black or white lines or dots.

Received 19 April; accepted 30 June 2004; doi:10.1038/nature02810.

- Bhatnagar, M. & Baliga, B. J. Comparison of 6H-SiC, 3C-SiC, and Si for power devices. *IEEE Trans. Electron Devices* **40**, 645–655 (1993).
- Harris, C. I., Savage, S., Konstantinov, A., Bakowski, M. & Ericsson, P. Progress towards SiC products. *Appl. Surf. Sci.* **184**, 393–398 (2001).
- Yakimova, R. *et al.* Seeded sublimation growth of 6H and 4H-SiC crystals. *Mater. Sci. Eng. B* **61–62**, 54–57 (1999).
- Selder, M., Kadinski, L., Durst, F. & Hofmann, D. Global modeling of the SiC sublimation growth process: prediction of thermoelastic stress and control of growth conditions. *J. Cryst. Growth* **226**, 501–510 (2001).
- Kato, T. *et al.* In-situ observation of silicon carbide sublimation growth by X-ray topography. *J. Cryst. Growth* **222**, 579–585 (2001).
- Oyanagi, N., Yamaguchi, H., Kato, T., Nishizawa, S. & Arai, K. Growth and evaluation of high quality SiC crystal by sublimation method. *Mater. Sci. Forum* **389–393**, 87–90 (2002).
- Frank, F. C. Capillary equilibria of dislocated crystals. *Acta Crystallogr.* **4**, 497–501 (1951).
- Heindl, J. *et al.* Dislocation content of micropipes in SiC. *Phys. Rev. Lett.* **80**, 740–741 (1998).
- Gutkin, M. Yu. *et al.* Synchrotron radiographic study and computer simulation of reactions between micropipes in silicon carbide. *J. Appl. Phys.* **94**, 7076–7082 (2003).
- Müller, St. G. *et al.* High quality SiC substrates for semiconductor devices: from research to industrial production. *Mater. Sci. Forum* **389–393**, 23–28 (2001).
- Müller, St. G. *et al.* Sublimation-grown semi-insulating SiC for high frequency devices. *Mater. Sci. Forum* **433–436**, 39–44 (2003).
- Neudeck, P. G., Huang, W. & Dudley, M. Breakdown degradation associated with elementary screw dislocations in 4H-SiC p+n junction rectifiers. *Solid-State Electron.* **42**, 2157–2164 (1998).
- Lendenmann, H. *et al.* Long term operation of 4.5 kV PiN and 2.5 kV JBS diodes. *Mater. Sci. Forum* **353–356**, 727–730 (2001).
- Malhan, R. K., Nakamura, H., Onda, S., Nakamura, D. & Hara, K. Impact of SiC structural defects on the degradation phenomenon of bipolar SiC devices. *Mater. Sci. Forum* **433–436**, 917–920 (2003).
- Senzaki, J., Kojima, K. & Fukuda, K. Long-term reliability of n-type 4H-SiC thermal oxides (3). In *Extended Abstracts of the 51st Spring Meeting of the JSAP* Vol. 1, 433 (The Japan Society for Applied Physics, Tokyo, 2004) [in Japanese].
- Tanimoto, S. *et al.* Impact of surface crystal-defects on TDD event of SiC thermal oxide. In *Extended Abstracts of the 51st Spring Meeting of the JSAP* Vol. 1, 434 (The Japan Society for Applied Physics, Tokyo, 2004) [in Japanese].
- Takahashi, J., Ohtani, N., Katsuno, M. & Shinoyama, S. Sublimation growth of 6H- and 4H-SiC single crystals in the [1100] and [1120] directions. *J. Cryst. Growth* **181**, 229–240 (1997).
- Dash, W. C. Growth of silicon crystals free from dislocations. *J. Appl. Phys.* **30**, 459–474 (1959).
- Zulehner, W. Historical overview of silicon crystal pulling development. *Mater. Sci. Eng. B* **73**, 7–15 (2000).
- Tairov, Y. M. & Tsvetkov, V. F. Investigation of growth processes of ingots of silicon carbide single crystals. *J. Cryst. Growth* **43**, 209–212 (1978).
- Matsunami, H. & Kimoto, T. Step-controlled epitaxial growth of SiC: High quality homoepitaxy. *Mater. Sci. Eng.* **R20**, 125–166 (1997).
- Stahlbush, R. E. *et al.* Propagation of current-induced stacking faults and forward voltage degradation in 4H-SiC PiN diodes. *Mater. Sci. Forum* **389–393**, 427–430 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank Y. Hirose for the SMBXT experiment, T. Saito for discussions, and N. Sugiyama, M. Matsui and H. Kuno for other experimental support. The SMBXT experiments were performed at the Spring-8 with the approval of the Japan Synchrotron Radiation Research Institute (JASRI).

Authors' contributions D.N. conceived the idea and the growth experiment, and together with I.G., A.O. and H.K. carried it out; D.N., S.Y. and T.I. executed the quality analysis; and D.N., S.O. and K.T. co-wrote the paper.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to K.T. (takatori@mosk.tytlabs.co.jp).

..... Ionic liquids and eutectic mixtures as solvent and template in synthesis of zeolite analogues

Emily R. Cooper, Christopher D. Andrews, Paul S. Wheatley, Paul B. Webb, Philip Wormald & Russell E. Morris

School of Chemistry, University of St Andrews, Purdie Building, St Andrews KY16 9ST, UK

The challenges associated with synthesizing porous materials¹ mean that new classes of zeolites (zeotypes)—such as aluminosilicate zeolites^{2,3} and zeolite analogues⁴—together with new methods of preparing known zeotypes⁵, continue to be of great importance. Normally these materials are prepared hydrothermally with water as the solvent in a sealed autoclave under autogenous pressure⁶. The reaction mixture usually includes an organic template or 'structure-directing agent' that guides the synthesis pathway towards particular structures. Here we report the preparation of aluminophosphate zeolite analogues by using ionic liquids⁷ and eutectic mixtures⁸. An imidazolium-based ionic liquid acts as both solvent and template, leading to four zeotype frameworks under different experimental conditions. The structural characteristics of the materials can be traced back to the solvent chemistry used. Because of the vanishingly low vapour pressure of ionic liquids, synthesis takes place at ambient pressure, eliminating safety concerns associated with high hydrothermal pressures. The ionic liquid can also be recycled for further use. A choline chloride/urea eutectic mixture⁸ is also used in the preparation of a new zeotype framework.

Ionic liquid solvents are commonly defined as salts that are fluid at near-ambient temperatures (less than $\sim 100^\circ\text{C}$)⁷ and consist of predominantly ionic species. For the purposes of this work we can use a broader definition of an ionic liquid as any salt that melts below the temperature used in the synthesis of zeolites (typically $150\text{--}200^\circ\text{C}$). Even when an organic salt melts at a higher temperature, mixing it with other compounds (such as urea or metal chlorides) can depress the melting point to produce liquids with significant ionic character at suitable temperatures. These are known as eutectic mixtures. A particular example of this is the hydroxyethyltrimethylammonium chloride (choline chloride, melting point $\sim 300^\circ\text{C}$) mixture with urea in a 1:2 ratio, which has a melting point of 12°C (ref. 8).

There are very few examples of the use of ionic liquids in the preparation of porous materials, and no reports of zeolites or microporous solids prepared by using ionic liquids or eutectic mixtures, either as solvents or templates. High-melting dialkylimidazolium hydroxide salts, which are not ionic liquids, have been used as templates in aqueous solvents⁹ and there is one report of the use of 1-alkyl 3-methyl imidazolium bromides to template a non-zeolitic porous solid, MCM-41, which is an amorphous silica material¹⁰. The pure organic salts in the latter example are ionic liquids. However, the actual solvent used in the preparation of MCM-41 is an aqueous solution containing only a small amount of the dialkylimidazolium salt. In this case the liquid phase is predominantly molecular, and the formation of MCM-41 relies on the water, which combines with the surfactant nature of the organic salt to produce the micelles required in the mechanism of the reaction. In both cases cited above the preparations use the traditional hydrothermal synthesis approach. However, ionic liquids have been used as solvents in the preparation of an aerogel¹¹. Here we present the first use of ionic liquids and eutectic mixtures in the preparation of crystalline zeolites. Not only do the ionic liquids act as solvent, they also provide the template cations around which the

inorganic frameworks order. We show further that the syntheses rely on the solvent's being predominantly ionic, because sufficient quantities of molecular water disrupt the reaction, preventing the formation of zeolites. We term this procedure ionothermal synthesis to distinguish it from hydrothermal preparations, which take place in a predominantly molecular solvent. In addition we show that reactant quantities of mineralizers such as fluoride can be used to control the product of the reactions.

1-Methyl 3-ethyl imidazolium bromide (melting point 83 °C) was used as both solvent and template in the synthesis of four different aluminophosphate zeotype frameworks, depending on the synthesis conditions used (Fig. 1). At 150 °C a novel structure type, SIZ-1 (St Andrews Ionic Liquid Zeotype-1) was formed, and its structure was solved by single-crystal X-ray diffraction on Station 9.8 at the Synchrotron Radiation Source, Daresbury, UK (see Supplementary Information for details). The structure of SIZ-1 (Fig. 2) consists of hexagonal prismatic units known as double six rings joined to form layers that are linked into a three-dimensional framework by units containing four tetrahedral centres (two phosphorus and two aluminium) known as single four rings. The formula of the material is $\text{Al}_8(\text{PO}_4)_{10}\text{H}_3.3\text{C}_6\text{H}_{11}\text{N}_2$ but the Al–O–P alternation is maintained. The framework is therefore interrupted, with

some unusual intraframework hydrogen bonding (Fig. 2). The negative charge present on the framework (caused by the existence of terminal P–O bonds¹²) balances the charge on the 1-methyl 3-ethyl imidazolium templates that are present in the pores. The overall structure of SIZ-1 (Fig. 2) shows a two-dimensional channel system parallel to the *a* and *b* crystallographic axes.

The choline chloride/urea eutectic mixture (melting point 12 °C) can also be used to prepare a novel zeotype framework, which we name SIZ-2. The Al–O–P alternation is maintained, but the chemical formula $(\text{Al}_2(\text{PO}_4)_3.3\text{NH}_4)$ indicates that SIZ-2 is also an interrupted structure (Fig. 3). In this case it is ammonium, formed on the partial decomposition of the urea, which acts to template the structure and balance the charge on the framework. The pore architecture displays three small intersecting channels parallel to the three crystallographic directions. These pores are too small to adsorb anything except the smallest molecules, but preliminary ion exchange experiments indicate that the ammonium can be at least partly exchanged for metal cations such as Cu^{2+} .

Interrupted aluminophosphate structures have the advantage of comprising anionic frameworks and can have interesting properties^{13–16}. The disadvantage of interrupted materials is that they are

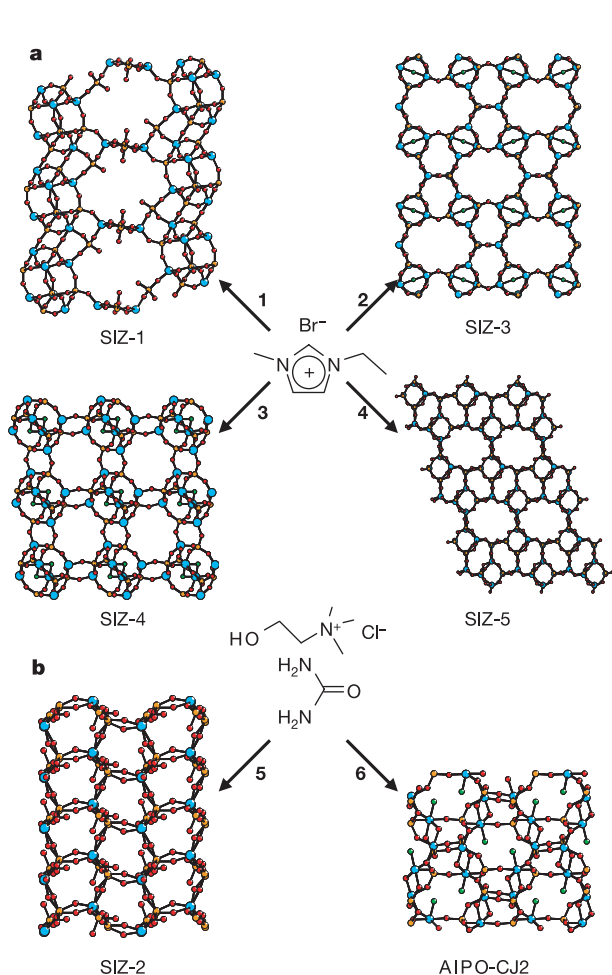


Figure 1 The synthesis of zeotypes by using ionic liquids and eutectic mixtures. **a**, 1-Ethyl 3-methyl imidazolium bromide can be used as both solvent and template to prepare SIZ-1, SIZ-3, SIZ-4 and SIZ-5. SIZ-3 and SIZ-4 are prepared in the presence of fluoride and SIZ-5 in the presence of excess water. **b**, A choline chloride/urea eutectic mixture can be used to prepare SIZ-2 in the absence of fluoride or excess water, and to prepare AIPO-CJ2 in the presence of fluoride or excess water. Conditions for reactions 1–6 are given in Table 1. Orange, cyan and red spheres correspond to phosphorus, aluminium and oxygen atoms respectively.

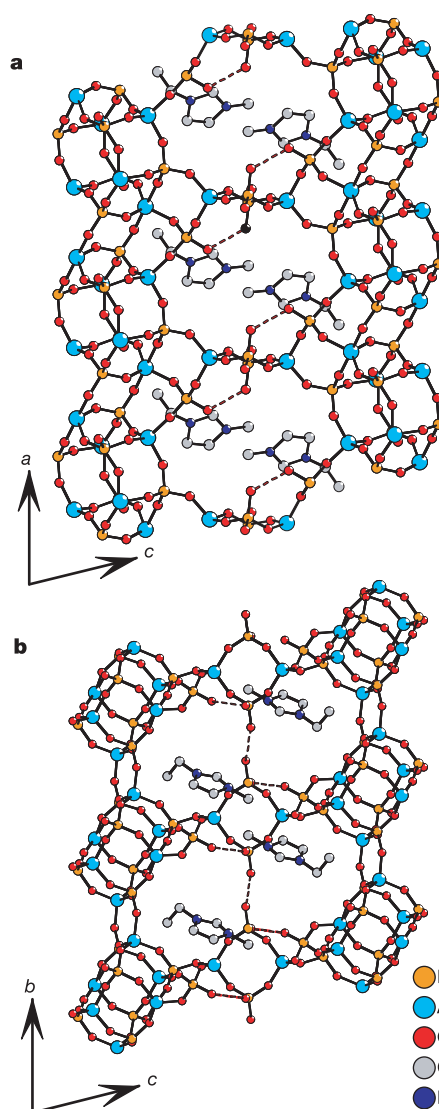


Figure 2 Two views of the SIZ-1 structure. **a**, Down the *b*-axis. **b**, Down the *a*-axis. In both views only one of the three disordered 1-ethyl 3-methyl imidazolium cations is shown for clarity. Intra-framework hydrogen bonds are shown as dotted lines.

invariably less thermally stable than their fully condensed counterparts. Mineralizing agents such as fluoride are known to promote T–O–T (where T represents a tetrahedral atom) bond formation¹⁷. Addition of fluoride to the synthesis mixtures does indeed lead to the formation of the fully condensed zeotype structure SIZ-3 ($\text{Al}_5\text{P}_5\text{O}_{20}\text{F}_2 \cdot 2\text{C}_6\text{H}_{11}\text{N}_2$), which has the AIPO-11 framework structure¹⁸. SIZ-4 ($\text{Al}_3\text{P}_3\text{O}_{12}\text{F} \cdot \text{C}_6\text{H}_{11}\text{N}_2$), which has the triclinic AIPO-34 structure^{19,20}, is also formed in the presence of fluoride but when

great care is taken to remove most of the water present in the starting materials by careful evaporation of the excess water in the reaction mixture before the reaction (Table 1). Addition of fluoride to the choline chloride/urea eutectic mixture results in the formation of a non-zeolitic aluminophosphate known as AIPO-CJ2 (Fig. 1)²¹. AIPO-CJ2 can also be prepared without fluoride but with excess water; hydroxide substitutes for fluoride in the final structure.

The lack of vapour pressure from the ionic liquid means that these synthetic procedures can be performed in open vessels, avoiding the high autogenous pressures (up to 15 bar at 200 °C) and associated safety concerns that accompany hydrothermal synthesis in sealed autoclaves²². SIZ-3 and SIZ-4 can both be prepared in round-bottomed flasks under ambient pressure by this method. In all cases the ionic liquid can be recycled and used in subsequent preparations.

The ionic liquid solubilizes the starting materials almost completely at the reaction temperatures, indicating that the synthesis mechanism is a crystallization from solution rather than a solid-to-solid transformation. The dependence of the products on water also gives some clues as to the mechanism of the reaction. Ionic liquids of this type are hydrophilic and it is almost impossible to dry them completely. However, infrared²³ and molecular dynamics studies²⁴ indicate that small amounts of water (for example at molar ratios of less than 10%) are present as isolated molecules that interact very strongly with the anions of the ionic liquids. The liquids in such situations are still overwhelmingly ionic in character. The interaction of water with the anions is probably so strong that it is effectively shielded from interaction with the Al- and P-containing ions that are solubilized by the cations of the ionic liquid. The interaction between the cations of the ionic liquid and the framework species (whether they are fully solubilized or at the surface of a growing crystallite) is the basis of the strong templating effect seen in these studies. The addition of fluoride ions helps to solubilize Al- and P-containing species in the ionic liquid even further through the formation of oxyfluoride anions in the solution, as well as catalysing the development of the Al–O–P bonds that are necessary for the crystallization of the solid¹⁷.

Potentially the most important feature of this ionothermal synthesis mechanism is the removal of the competition between template–framework and solvent–framework interaction that is present in hydrothermal preparations. The structure-directing properties of templates are often not as specific as we should like²⁵, but in a system in which an ionic liquid is both solvent and template the negatively charged atoms at the surface of a growing framework will always be interacting primarily with the templating cation rather than with a mixture of template and solvent. Recent modelling studies indicate that the structure of imidazolium-based ionic liquids is characterized by long-range correlations and distributions that reflect the asymmetric structures of the cations²⁶. Long-range asymmetric effects of this kind potentially increase the like-

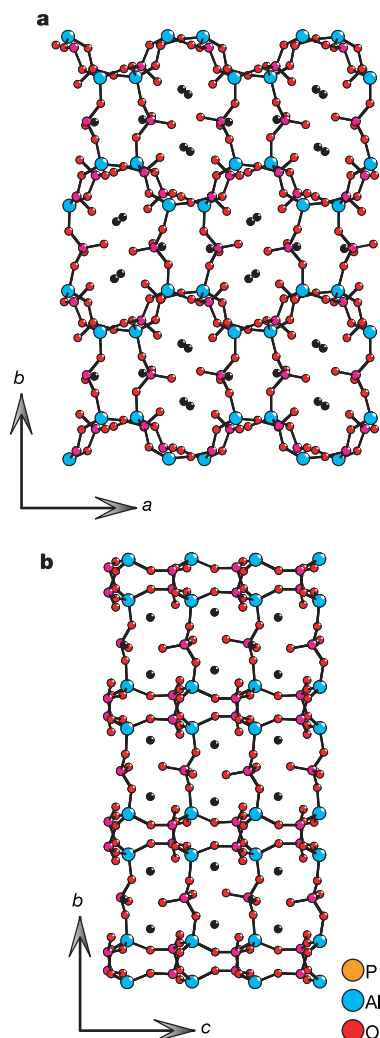


Figure 3 Two views of the structure of SIZ-2. **a**, Down the *c*-axis. **b**, Down the *a*-axis. The nitrogen atoms from the ammonium ions located in the pores of the structure are shown as black spheres.

Table 1 Synthesis details and conditions for the preparation of materials

Product	Mass of reagents added (g) (molar ratio of reagents)				IL	Temp. (°C)	Time (h)
	$\text{Al}(\text{O}i\text{Pr})_3$	H_3PO_4	HF	Water			
SIZ-1	0.1018 (1.0)	0.1732 (3.0)	0.00 (0.0)	* (2.9)	4.05 (43)	150	66
SIZ-3	0.1013 (1.0)	0.1732 (3.0)	0.015 (0.73)	* (3.8)	4.07 (43)	150	68
SIZ-4	0.1054 (1.0)	0.1772 (3.0)	0.015 (0.70)	* (0.0)	3.82 (39)	150	68
SIZ-5	0.0465 (1.0)	0.0858 (3.1)	0.00 (0.0)	0.495 (116)	2.03 (44)	150	19
Product	Mass of reagents added (g) (molar ratio of reagents)				EU	Temp. (°C)	Time (h)
	$\text{Al}(\text{O}i\text{Pr})_3$	H_3PO_4	HF	Water			
SIZ-2	0.1007 (1.0)	0.171 (3.0)	0.00 (0.0)	* (2.9)	5.13 (40)	180	72
ALPO-CJ2†	0.1059 (1.0)	0.175 (2.9)	0.00 (0.0)	0.557 (62)	4.91 (37)	180	72
ALPO-CJ2‡	0.1024 (1.0)	0.178 (3.1)	0.015 (0.7)	* (3.8)	5.32 (41)	180	72

$\text{Al}(\text{O}i\text{Pr})_3$, $\text{Al}[\text{OCH}(\text{CH}_3)_2]_3$; EU, eutectic mixture; IL, ionic liquid.

*No extra water was added in these preparations. Small amounts of water present came from the aqueous HF and H_3PO_4 solutions and the ionic liquid.

†No added fluoride.

‡Fluoride added.

likelihood of transferring chemical information from the template cation to the framework, a situation that is desirable if full control over the templating process is to be achieved.

With the addition of larger quantities of water the liquid becomes less ionic and more molecular in character, and water molecules begin to form large hydrogen-bonded clusters and networks that percolate through the whole liquid phase²⁴. This change in the molecular structure of the solvent changes the reactions drastically, and the dense phase berlinite (AlPO₄) becomes the favoured product. The presence of a hydrogen-bonded water network thus disrupts the ionic-liquid-framework interactions to such a degree that they are completely removed from the important crystallization steps in the reaction mechanism. As even more water is added so that it becomes by far the major chemical species, the solvent becomes predominantly molecular and cannot now be described as an ionic liquid. The reaction in such solvents is hydrothermal in nature and is similar to the traditional method of preparing zeolites.

The trends in reaction products are the same for both the fluoride and non-fluoride syntheses. Under ionothermal conditions (low or zero added water) the zeotypes described above are produced. At intermediate water concentrations only dense phases are produced and under hydrothermal conditions a different zeotype framework is formed; at water contents in which the ionic liquid/water ratio is less than 1:2.5, SIZ-5—which has the AlPO-41 framework²⁷—is the major product, with a small amount of dense-phase impurity (Table 1).

Functional aluminophosphates require aliovalent doping to produce an anionic framework. Typically this is achieved by using silicon to produce silicoaluminophosphate materials. Silicon can be incorporated into the SIZ-4 structure with the use of this synthetic procedure, pointing to potential catalytic applications of such materials. Gallium can also be doped into the structure in the same way.

The use of ionic liquids and eutectic mixtures as solvent and template opens up many new possibilities in the preparation of zeolites and related porous solids. The relative lack of mineralizers in the pure ionic liquid solvent is probably the controlling reason why interrupted structures are formed when no fluoride is added. This gives us the possibility of selectively targeting interrupted frameworks (for example, for ion-exchange applications) or fully condensed zeotypes depending on which type of material is required. The scope for this preparative method is further emphasized if one realizes that there are at least a million binary ionic liquids and potentially more than 10¹⁸ ternary ionic liquids and eutectic mixtures⁷ (compared with only about 600 molecular solvents, most of which have a polarity that makes them completely unsuited to this type of chemistry). Many of these ionic liquids will be structurally related to each other and might template the same frameworks (so-called default structures), so it is unlikely that 10⁶ new materials will be prepared. However, the results reported here indicate that multiple frameworks are possible even from the same ionic liquid system. Computational enumeration predicts that several thousand zeolite structures are possible²⁸, and there is clearly scope for new synthesis methodologies to target these new materials. As long as kinetic considerations do not limit the method to the production of a small number of default structures, the ionothermal approach reported here might be a route for the preparation of many previously unknown materials. We are currently extending the ionic liquid chemistry to explore whether this is a general route to novel solids and to establish the chemical properties of the materials prepared. □

Methods

Synthesis of 1 ethyl 3-methyl imidazolium bromide

1-Ethyl 3-methyl imidazolium bromide (hereafter referred to as IL) was prepared in 94% yield from 1-methylimidazole and ethyl bromide as described²⁹.

Synthesis of choline chloride/urea eutectic mixture

Choline chloride/urea eutectic mixture was prepared from choline chloride and urea in a 1:2 ratio as described⁸.

Synthesis of zeolite analogues in sealed autoclaves

A typical synthesis procedure was as follows: a Teflon-lined autoclave (volume 23 ml) was charged with IL or choline chloride/urea eutectic mixture, Al[OCH(CH₃)₂]₃ (Aldrich) and H₃PO₄ (85 wt% in water; Aldrich). Distilled water or HF (48 wt% in water; Aldrich) was added if required. The stainless steel autoclave was then heated in an oven to the required temperature. The reagent masses, temperatures and length of time left in oven needed to produce the pure phase materials are as detailed in Table 1. These conditions were optimized by changing the reaction compositions (for example, the water content) slightly and characterizing the resulting products. The yields for the reactions, quoted with respect to the aluminium-containing starting material, range from 50–60% for SIZ-1 (51%), SIZ-3 (56%) and SIZ-4 (50%) to almost 100% for SIZ-1 and SIZ-5. However, it should be noted that both SIZ-1 and SIZ-5 samples tend to contain small amounts of amorphous phosphate-rich material.

For the SIZ-4 synthesis the Al[OCH(CH₃)₂]₃, H₃PO₄ and HF were added to the Teflon liner. This was then heated to 50 °C on a hotplate for 2 h to remove the water and any propan-2-ol formed during the initial reaction. This was confirmed by following the mass of the mixture until no more mass loss was observed. The IL was then added and the synthesis proceeded as normal.

After the autoclave had been cooled to room temperature the product was suspended in distilled water, filtered by suction and washed with acetone. The products were all white crystalline solids.

Replacement of the IL with the symmetric ionic liquid (1,1'-dimethyl 3,3'-hexamethylene diimidazolium dibromide) also leads to the production of SIZ-3 or SIZ-4 under the same conditions as those in Table 1. Silicon can be incorporated into the SIZ-4 framework structure by adding about 0.06 mol equivalents of a silica source (Cab-O-Sil) into the reaction listed in Table 1. Similarly gallium can be incorporated by adding Ga₂SO₄ to the reaction mixtures.

¹³C magic-angle spinning NMR indicates that the 1-methyl 3-ethyl imidazolium cation is intact in the pores in SIZ-1, SIZ-3, SIZ-4 and SIZ-5; its presence is confirmed by single-crystal X-ray diffraction studies in SIZ-1 and SIZ-4. SIZ-3, SIZ-4 and SIZ-5 are all very thermally stable. Heating in oxygen at 500 °C leads to the removal of the organic cation and the fluoride to leave the empty zeotype frameworks. The interrupted structures, SIZ-1 and SIZ-2, are less thermally stable and collapse at ~250 °C as the organics are removed from the structures (see Supplementary Information for full characterization details of all materials).

Synthesis of zeolite analogues in round-bottomed flasks

SIZ-3 and SIZ-4 can be prepared in an open container using the conditions from Table 1. A round-bottomed flask fitted with a condenser, magnetic stirrer and drying tube was charged with the starting materials and then heated to the target temperature (150 °C) for the required duration. The products were recovered as described above.

Recycling of ionic liquid

After the zeotype had been filtered off, the filtrate was centrifuged to remove any remaining small solid particles. The water and acetone were removed from the ionic liquid by rotary evaporation. The remaining product was dissolved in excess dichloromethane and magnesium sulphate was added. This was stirred for 30 min and then filtered under suction. The dichloromethane was removed by rotary evaporation. Ethyl acetate was added and the solidified ionic liquid precipitated from solution. This was filtered, then washed with ethyl acetate and dried under vacuum at 25 °C for 10 h to give IL. About 80% of the mass of ionic liquid used in the original preparation could be recovered. Note, however, that some was used up as a template and so was not available for recycling; the real recycling efficiency was therefore greater than 80%. The ionic liquid was characterized by ¹H NMR. The IL was then used successfully in the preparation of zeotypes SIZ-3 and SIZ-4.

Received 21 April; accepted 16 July 2004; doi:10.1038/nature02860.

1. Davis, M. E. Ordered porous materials for emerging applications. *Nature* **417**, 813–821 (2002).
2. Corma, A., Díaz-Cabañas, M. J., Joaquín Martínez-Triguero, J., Rey, F. & Rius, J. A large-cavity zeolite with wide pore windows and potential as an oil refining catalyst. *Nature* **418**, 514–517 (2002).
3. Freyhardt, C. C., Tsapatsis, M., Lobo, R. F., Balkus, K. J. Jr & Davis, M. E. A high-silica zeolite with a 14-tetrahedral-atom pore opening. *Nature* **381**, 295–298 (1996).
4. Feng, P., Bu, X. & Stucky, G. D. Hydrothermal syntheses and structural characterization of zeolite analogue compounds based on cobalt phosphate. *Nature* **388**, 735–741 (1997).
5. Lee, H., Zones, S. I. & Davis, M. E. A combustion-free methodology for synthesizing zeolites and zeolite-like materials. *Nature* **425**, 385–388 (2003).
6. Cundy, C. S. & Cox, P. A. The hydrothermal synthesis of zeolites: history and development from the earliest days to the present time. *Chem. Rev.* **103**, 663–701 (2003).
7. Rogers, R. D. & Seddon, K. R. Ionic liquids—solvents of the future? *Science* **302**, 792–793 (2003).
8. Abbott, A. P., Capper, G., Davies, D. L., Rasheed, R. K. & Tambyrajah, V. Novel solvent properties of choline chloride/urea mixtures. *Chem. Commun.* 70–71 (2003).
9. Zones, S. I. Synthesis of pentasil zeolites from sodium-silicate solutions in the presence of quaternary imidazole compounds. *Zeolites* **9**, 458–467 (1989).
10. Adams, C. J., Bradley, A. E. & Seddon, K. R. The synthesis of mesoporous materials using novel ionic liquid templates in water. *Aust. J. Chem.* **54**, 679–681 (2001).
11. Dai, S. *et al.* Preparation of silica aerogel using ionic liquids as solvents. *Chem. Commun.* 243–244 (2000).
12. Yu, J. H. & Xu, R. R. Rich structure chemistry in the aluminophosphate family. *Acc. Chem. Res.* **36**, 481–490 (2003).

13. Eloisa Medina, M., Iglesias, M., Gutiérrez-Puebla, E. & Angeles Monge, M. Solvothermal synthesis and structural relations among three anionic aluminophosphates; catalytic behaviour. *J. Mater. Chem.* **14**, 845–850 (2004).
14. Huo, Q. *et al.* Synthesis and characterization of a novel extra large ring of aluminophosphate JDF-20. *Chem. Commun.* 875–876 (1992).
15. Yu, J. *et al.* $\text{Al}_{16}\text{P}_{20}\text{O}_{80}\text{H}_4\text{C}_6\text{H}_{18}\text{N}_2$: A new microporous aluminophosphate containing intersecting 12- and 8-membered ring channels. *Chem. Mater.* **10**, 1208–1211 (1998).
16. Estermann, M., McCusker, L. B., Baerlocher, C., Merrouche, A. & Kessler, H. A synthetic gallophosphate molecular-sieve with a 20-tetrahedral-atom pore opening. *Nature* **352**, 320–323 (1991).
17. Cambor, M. A., Villaescusa, L. A. & Diaz-Cabanas, M. J. Synthesis of all-silica and high-silica molecular sieves in fluoride media. *Top. Catal.* **9**, 59–76 (1999).
18. Richardson, J. W. Jr, Pluth, J. J. & Smith, J. V. Rietveld profile analysis of calcined AlPO_4 -11 using pulsed neutron powder diffraction. *Acta Crystallogr. B* **44**, 367–373 (1988).
19. Oliver, S., Kuperman, A., Lough, A. & Ozin, G. A. Synthesis and characterization of a fluorinated anionic aluminophosphate framework UT-6, and its high-temperature dehydrofluorination to AlPO_4 -CHA. *J. Mater. Chem.* **7**, 807–812 (1997).
20. Feng, P., Bu, X., Gier, T. E. & Stucky, G. D. Amine-directed syntheses and crystal structures of phosphate-based zeolite analogs. *Microporous Mesoporous Mater.* **23**, 221–229 (1998).
21. Taulelle, F. *et al.* Isomerization of the prenucleation building unit during crystallization of AlPO_4 -CJ2: An MQMAS, CP-MQMAS, and HETCOR NMR study. *J. Am. Chem. Soc.* **121**, 12148–12153 (1999).
22. Robson, H. in *Verified Synthesis of Zeolitic Materials* (ed. Robson, H.) 45–46 (Elsevier, Amsterdam, 2001).
23. Cammarata, L., Kazarian, S. G., Salter, P. A. & Welton, T. Molecular states of water in room temperature ionic liquids. *Phys. Chem. Chem. Phys.* **3**, 5192–5200 (2001).
24. Hanke, C. G. & Lynden-Bell, R. M. A simulation study of water-dialkylimidazolium ionic liquid mixtures. *J. Phys. Chem. B* **107**, 10873–10878 (2003).
25. Lobo, R. F., Zones, S. I. & Davis, M. E. Structure-direction in zeolite synthesis. *J. Inclusion Phenom. Mol.* **21**, 47–78 (1995).
26. Del Popolo, M. G. & Voth, G. A. On the structure and dynamics of ionic liquids. *J. Phys. Chem. B* **108**, 1744–1752 (2004).
27. Kirchner, R. M. & Bennett, J. M. The structure of calcined AlPO_4 -41: A new framework topology containing one-dimensional 10-ring pores. *Zeolites* **14**, 523–528 (1994).
28. Delgado-Friedrichs, O., Dress, A. W. M., Huson, D. H., Klinowski, J. & Mackay, A. L. Systematic enumeration of crystalline networks. *Nature* **400**, 644–657 (1999).
29. Wasserscheid, P. & Welton, T. *Ionic Liquids in Synthesis* Ch. 2.1.2, 9–12 (Wiley-VCH, Weinheim, Germany, 2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Hardacre (Queen's University, Belfast) and D. Cole-Hamilton (University of St Andrews) for discussions. We thank the CCLRC and S. Teat for access to the Synchrotron Radiation Source (Daresbury) and the EPSRC and D. Apperley for access to solid-state NMR. R.E.M. was supported by a Royal Society University Research Fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.E.M. (rem1@st-and.ac.uk).

Similar meltwater contributions to glacial sea level changes from Antarctic and northern ice sheets

Elco J. Rohling¹, Robert Marsh¹, Neil C. Wells¹, Mark Siddall^{1,2} & Neil R. Edwards^{2,3}

¹Southampton Oceanography Centre, Southampton SO14 3ZH, UK

²Climate and Environmental Physics, University of Bern, Sidlerstrasse 5, and

³NCCR-Climate, University of Bern, Erlachstrasse 9a, CH-3012 Bern, Switzerland

The period between 75,000 and 20,000 years ago was characterized by high variability in climate^{1–12} and sea level^{13,14}. Southern Ocean records of ice-rafted debris¹⁵ suggest a significant contribution to the sea level changes from melt water of Antarctic origin, in addition to likely contributions from northern ice sheets, but the relative volumes of melt water from northern and southern sources have yet to be established. Here we simulate the

first-order impact of a range of relative meltwater releases from the two polar regions on the distribution of marine oxygen isotopes, using an intermediate complexity model. By comparing our simulations with oxygen isotope data from sediment cores, we infer that the contributions from Antarctica and the northern ice sheets to the documented sea level rises between 65,000 and 35,000 years ago¹³ were approximately equal, each accounting for a rise of about 15 m. The reductions in Antarctic ice volume implied by our analysis are comparable to that inferred previously for the Antarctic contribution to meltwater pulse 1A (refs 16, 17), which occurred about 14,200 years ago, during the last deglaciation.

Greenland ice-core records show strong climate fluctuations between 75 and 20 kyr BP: the Dansgaard-Oeschger (DO) cycles^{1–3}. Abrupt warmings of 6–10 °C occurred throughout the mid- to high-latitude North Atlantic region at a spacing of about 1,500 yr (refs 4, 5). Each initiated a relatively warm DO interstadial, during which a gradual cooling trend developed that eventually culminated in rapid 'collapse' to the next cold DO stadial. DO-style variability was widespread throughout the northern hemisphere and beyond^{5–9}.

Antarctic ice-core records follow a different 'rhythm', with fewer and temporally more symmetrical climate fluctuations. The timing relationship between DO-style and Antarctic-style fluctuations was established with the use of atmospheric methane data from ice cores¹⁰ and was corroborated by work on marine sediment core MD952042 from 3,146 m depth off Portugal¹². The $\delta^{18}\text{O}$ record for surface-water planktonic foraminifera in MD952042 shows DO-style variability, whereas that for bottom-dwelling benthic foraminifera mimics Antarctic-style variability¹². The latter is observed also in SW Pacific core MD972120 from 1,210 m depth⁹, showing the global nature of the benthic signal.

The widespread benthic $\delta^{18}\text{O}$ signal suggests a relationship between variations in southern high-latitude climate and global ice volume (sea level). However, benthic $\delta^{18}\text{O}$ -based sea level reconstructions might be biased by deep-sea temperature changes: 1 °C error translates to about 30 m uncertainty in sea level. Efforts to separate ice volume from temperature influences on benthic $\delta^{18}\text{O}$ indicate rapid and high-amplitude ice-volume variability during the last glacial cycle¹⁴. Independent sea level quantification from the Red Sea method is coherent with both the absolute values from fossil reef data and the structure of benthic $\delta^{18}\text{O}$ records¹³. It indicates that sea level rose by about 30 m, at 2 m per century, in association with the 2–3 °C warming of Antarctic climate events A1–4 (Fig. 1a)¹³.

Although the Antarctic-style timing of sea level change might imply Antarctic ice-volume variations, it is equally possible that southern high-latitude climate fluctuations were driven by oscillations in climate and ice volume on the Northern Hemisphere^{18,19}, because Antarctic ice volume is often considered relatively stable, with its full glacial–interglacial variability contributing less than 25 m to the roughly 120 m global sea level change^{20–24}, and because models forced with meltwater additions into the North Atlantic and Arctic show interhemispheric temperature fluctuations similar to those observed in ice cores^{18,19}. However, recent work has challenged the notion of Antarctic stability on the basis of distinct ice-rafted debris (IRD) peaks in Southern Ocean records, preceded by –0.5‰ to –0.9‰ shifts in surface-water foraminiferal $\delta^{18}\text{O}$ (ref. 15; Fig. 1c). A predominantly Antarctic origin has also been inferred for meltwater pulse (mwp) 1A, a sea level rise of about 20 m in about 500 yr during the last deglaciation^{16,17}.

Here we assess the origins of the meltwater pulses that caused sea level rises associated with Antarctic warming events A1–4 (Fig. 1a), by using marine $\delta^{18}\text{O}$ as a sensitive tracer for input of isotopically light high-latitude melt water. Away from its surface-bound source and sink terms, $\delta^{18}\text{O}$ in sea water behaves as a conservative, passive tracer. Crucially, past seawater $\delta^{18}\text{O}$ changes are reflected in $\delta^{18}\text{O}$

records from fossil foraminiferal shells (carbonate) in sediment cores, the only significant complication being a temperature-dependent $\delta^{18}\text{O}_{\text{water}}$ to $\delta^{18}\text{O}_{\text{carbonate}}$ fractionation (1‰ shift to lighter $\delta^{18}\text{O}_c$ values for every 4°C warming). Bearing this in mind, we use fossil records to validate $\delta^{18}\text{O}$ distribution simulations that we develop using an Earth System Model of Intermediate Complexity (EMIC), namely the three-dimensional ocean/atmosphere/sea-ice model C-GOLDSTEIN^{25,26} (Methods). To assess the impacts of meltwater influxes at typical magnitudes and rates (Fig. 1), we impose a total release of 345 Sv (1 Sv = $10^6 \text{ m}^3 \text{ s}^{-1}$) years (30 m sea level rise over 1,500 yr). Robustness of results is

assessed through sensitivity analyses (Methods), and maps and profiles are given for a variety of scenarios in Supplementary Information. We initially focus on $\delta^{18}\text{O}_w$ (Figs 2 and 3a, b) but also consider $\delta^{18}\text{O}_c$ maps and profiles that rely on modelled temperatures, based on $\Delta\delta^{18}\text{O}_c = \Delta\delta^{18}\text{O}_w - (1/4)\Delta T$ (Fig. 3b–d).

Model simulations are validated with foraminiferal $\delta^{18}\text{O}_c$ records from the wider Atlantic Ocean (equivalent to the Arctic plus the North and South Atlantic) through the intervals of the A1–4 warming events, with emphasis on the upper 2,000 m, where the most distinctive signal amplitudes are found. Key $\delta^{18}\text{O}_c$ observations for deep-water (benthic foraminifera) and surface-

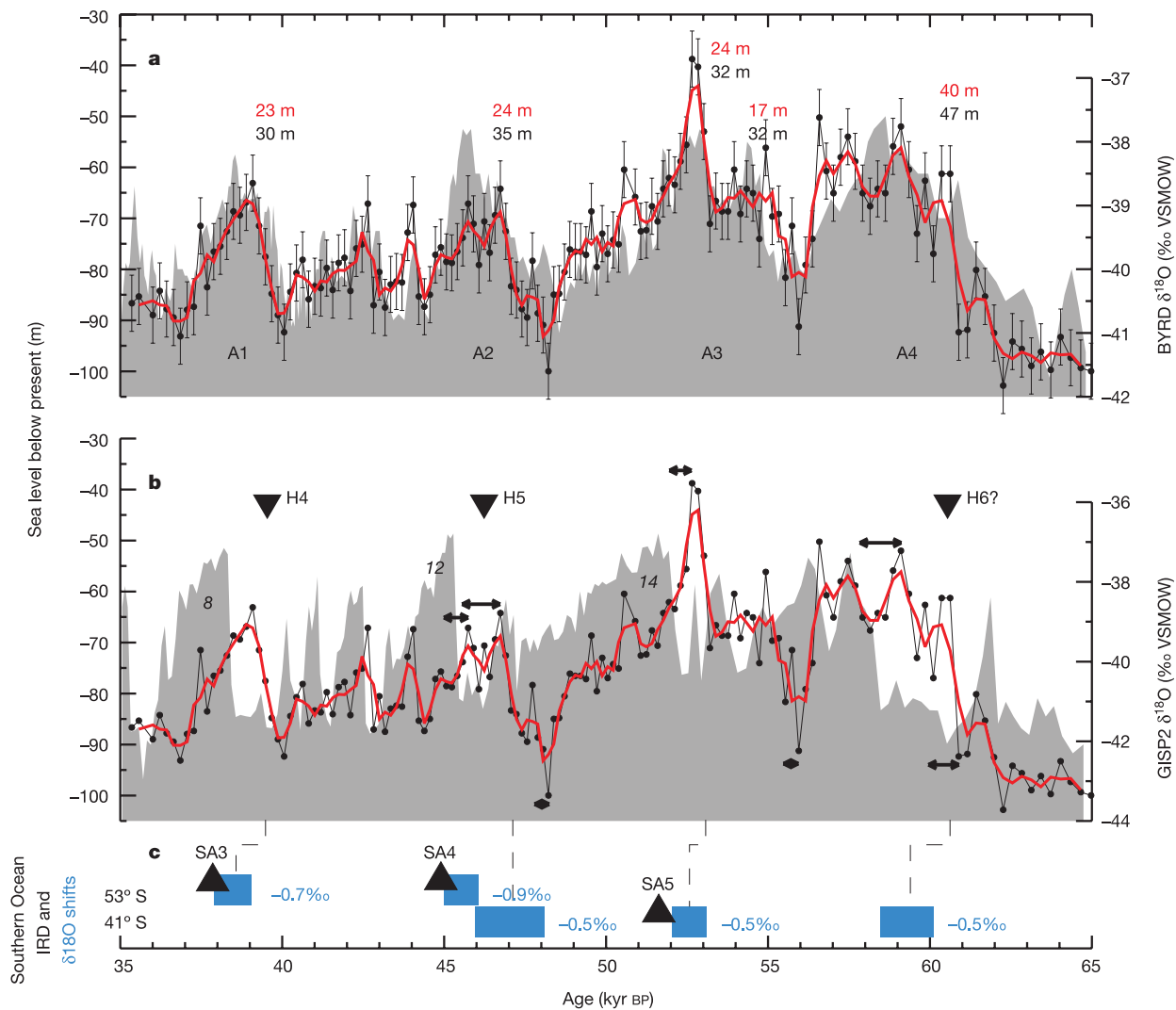


Figure 1 Comparison of the Red Sea sea level record with Antarctic and Greenland ice-core records and with the timing of both North Atlantic (Heinrich; H) and Southern Ocean (SA) IRD events. **a**, Sea level reconstruction from the Red Sea method (black dots and line)¹³ compared with Antarctic ice-core $\delta^{18}\text{O}$ temperature proxy data (grey) from the BYRD-station ice core (80.0° S, 119.5° W). Error bars indicate 1 standard deviation. **b**, As **a**, but for Greenland Ice Sheet Project 2 (GISP2, 72.6° N, 38.5° W) ice-core $\delta^{18}\text{O}$ (grey). Ice-core data as synchronized to the GISP2 timescale using variations in atmospheric methane concentrations¹⁰. Red lines represent three-point (about 500-yr) moving averages of the sea level record, which smoothes inter-sample variability within bounds of the 1σ interval (± 5.5 m) that applies to the sea level reconstruction¹³. Red numbers in **a** indicate magnitudes of increases in sea level according to the smoothed record, whereas black numbers relate to the unsmoothed data. Inverted triangles in **b** indicate the occurrence of North Atlantic Heinrich Events H4, H5 and tentatively H6, relative to

the GISP2 record⁸. Italic numbers identify main Dansgaard–Oeschger interstadials 8, 12 and 14. Double-headed arrows illustrate dating differences¹³ between the original correlation of the Red Sea sea level record to BYRD (as in **a**) and after tuning to a benthic $\delta^{18}\text{O}$ record¹². **c**, Schematic representation of intervals of IRD, marked by triangles and SA numbers, as well as intervals with negative $\delta^{18}\text{O}$ shifts in planktonic foraminiferal calcite at 41° and 53° S in the South Atlantic sector¹⁵. These events were placed within our time frame by graphically transferring the positions of the events relative to the GISP2 $\delta^{18}\text{O}$ record as originally reported¹⁵. Dashed lines indicate possible correlations to the sea level record. A maximum disagreement is suggested of roughly 1,000 yr, which seems reasonable given the inherent uncertainties in both the sea level chronology (**b**, and ref. 13) and Southern Ocean records¹⁵. VSMOW, Vienna Standard Mean Ocean Water.

water (planktonic foraminifera) are reviewed in Supplementary Information, and appropriate ranges for the observations are represented with coloured circles in Fig. 3b–d.

For 100% northern meltwater release, the model shows a collapse

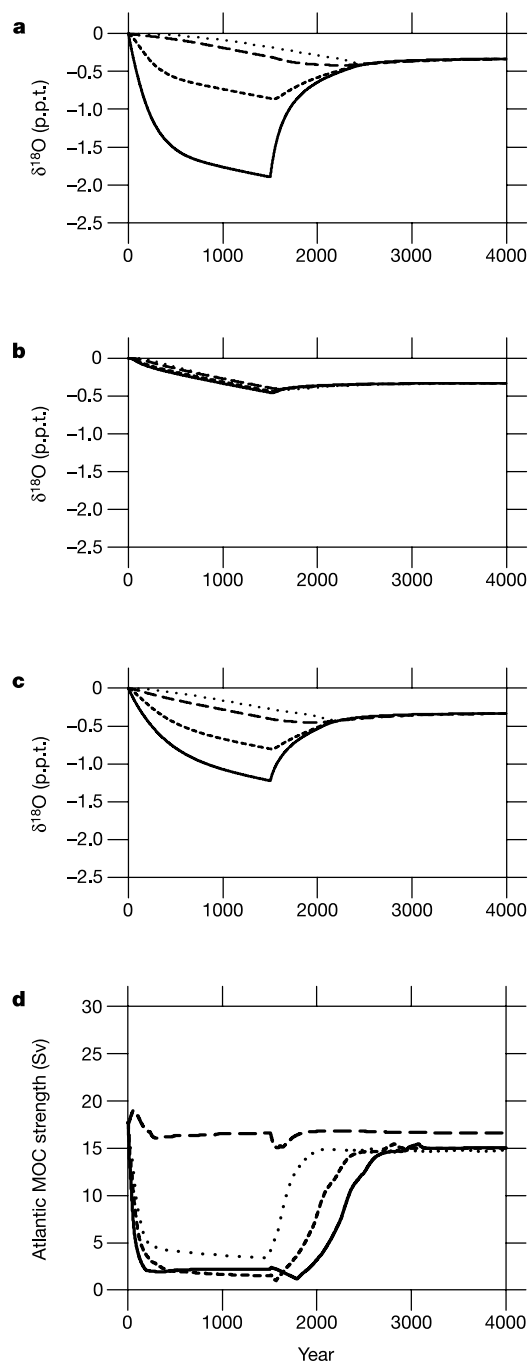


Figure 2 Development of Atlantic-wide mean $\delta^{18}\text{O}$ values with time in the upper (solid lines), intermediate (broken lines), deep (dashed lines) and bottom (dotted lines) waters. All simulations performed with the EMIC concern a total meltwater release of 0.23 Sv over 1,500 yr. **a–c**, Three scenarios: 100% release into the North Atlantic (**a**), 100% release around Antarctica (**b**), and 50% release into the North Atlantic and 50% around Antarctica (**c**). **d**, Changes in the MOC intensity for the various scenarios, including a 30:70 experiment that closely approximates the collapse of the Atlantic MOC. More detailed graphs of results are given in Supplementary Information. Solid line, 100% release into the Atlantic; broken line, 50% release into the North Atlantic and 50% around Antarctica; dotted line, 30% release into the North Atlantic and 70% around Antarctica; dashed line, 100% release around Antarctica.

of the Atlantic meridional overturning circulation (MOC) (Fig. 2d). This ‘traps’ the meltwater influence in the top 1,200 m of the Atlantic, and particularly the top 400 m, so that the Atlantic-wide mean $\delta^{18}\text{O}_w$ anomaly in the upper 400 m reaches -1.9‰ , whereas the mean value between 400 and 1,200 m is -0.9‰ (Fig. 2a). The most extreme $\delta^{18}\text{O}_c$ values produced by this scenario in the surface North Atlantic are of the order of -5‰ . This simulated range of values is incompatible with the observations (Fig. 3c).

For 100% meltwater release around Antarctica, the Atlantic MOC intensifies to about 19 Sv within about 100 yr, followed by gradual weakening over the next couple of centuries (Fig. 2d). There is no pooling of melt water at the surface in the North Atlantic, and $\delta^{18}\text{O}_w$ anomalies do not exceed -0.5‰ (Fig. 2b and Supplementary Information). This scenario fails to capture the considerable regional $\delta^{18}\text{O}_c$ anomalies that have been observed in the North Atlantic¹¹ (Supplementary Information), and also produces deep-water $\delta^{18}\text{O}_c$ fields that are very different from the observations (Fig. 3d).

Meltwater discharge into the NW Atlantic during Heinrich events of IRD deposition (Fig. 1) may have caused a global sea level rise of up to 15 m within 500 ± 250 yr (ref. 11). Our next experiment therefore considered 50% release into the North Atlantic and 50% around Antarctica. In the absence of clearly documented phase relationships between northern and southern meltwater releases, we specify both as synchronous and lasting 1,500 yr. Different phasing and durations were considered separately (below).

The 50:50 scenario shows a collapse of the Atlantic MOC (Fig. 2d), causing an Atlantic-wide mean $\delta^{18}\text{O}_w$ anomaly of -1.2‰ in the upper 400 m, and -0.8‰ between 400 and 1,200 m (Figs 2c and 3a, b). The most extreme North Atlantic $\delta^{18}\text{O}_c$ values approach -3‰ . Cooling in much of the North Atlantic surface to intermediate waters (Fig. 4c, d) would reduce the amplitude of the simulated $\delta^{18}\text{O}$ anomalies when recorded in carbonates. Maps and profiles of $\delta^{18}\text{O}_c$ show expected anomalies in broad agreement with the observations from sediment cores, although it seems that the model somewhat underestimates cooling to the north of about 60°N (Fig. 3b).

Further experiments at 10% increment changes between the northern and southern inputs corroborate the notion that agreement with the isotope data constrains the meltwater contribution from both regions to roughly equal ($50 \pm 10\%$) proportions. This first-order conclusion of roughly equal contributions is not significantly affected by changes in the $\delta^{18}\text{O}$ values of the northern and southern meltwater components to -30‰ and -50‰ , respectively (Supplementary Information).

How well does the 50:50 scenario simulate the bipolar temperature seesaw inferred from Greenland and Antarctic ice cores¹⁰? It shows robust cooling in the north and warming in the south (Fig 4a, b), but the simulated southern warming underestimates the $2\text{--}3^\circ\text{C}$ deduced from ice cores²⁷. Note, however, that all experiments concern highly schematic scenarios in a model of intermediate complexity. Inclusion of the radiative effects of a simple 20 p.p.m.v. CO_2 fluctuation²⁸, for example, brings the modelled temperature amplitudes closer to observed values (Fig. 4e, f), indicating that future models should include a reasonable carbon cycle. Separate tests with offset phasing and different durations of the northern and southern meltwater releases highlight a dominant control on the temporal structure of the temperature signals, making these parameters prime targets for future data collection.

In dynamical terms, our experiments simulate a ‘realistic’ range of isotope values when we apply sufficient meltwater release into the North Atlantic to severely reduce or just collapse the Atlantic MOC. To approximate the bipolar temperature seesaw also, a collapse seems to be required. The model shows a robust collapse when about one-third of the stipulated meltwater release is specified through any route into the North Atlantic (either at high latitudes or through the Gulf of Mexico) (Fig. 2d, and Methods), similar

For the first time we find a meltwater scenario that reasonably approximates the main characteristics of the sea level, marine $\delta^{18}\text{O}$, and seesaw data. Our robust first-order conclusion is that the investigated increases in sea level (Fig. 1) resulted from both

meltwater pulses are comparable to those suggested for mwp 1A (refs 16, 17). We find that such events occurred at a 5–8-kyr spacing between 60 and 35 kyr BP. Three events peaked about 1,000 yr before major DO interstadials (Fig. 1b). This is sufficiently similar to the

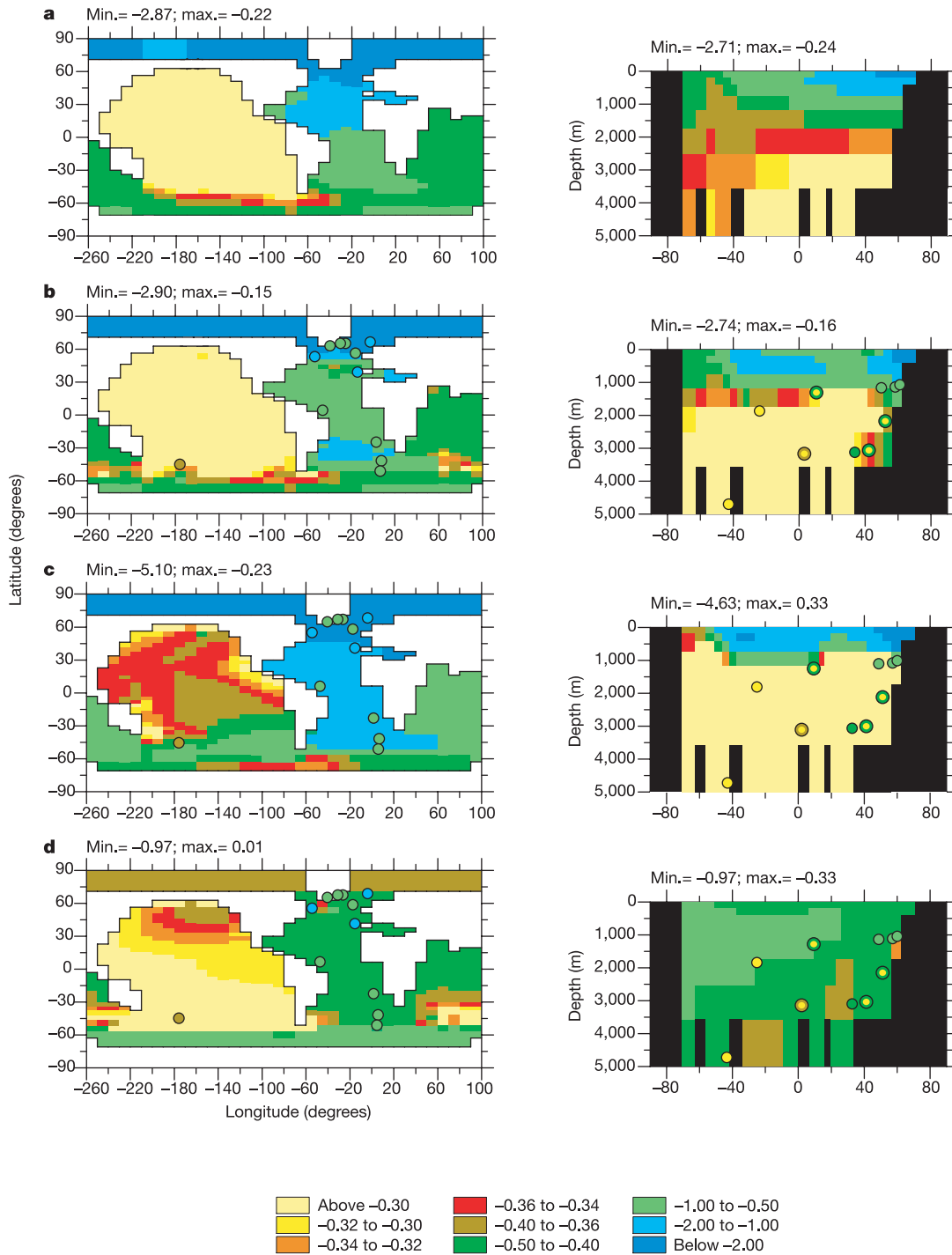


Figure 3 Horizontal (at 80 m depth) and vertical (along 25° W) distributions of $\delta^{18}\text{O}$ after 1,500 yr of 0.23 Sv meltwater release. **a, b**, $\delta^{18}\text{O}_{\text{w}}$ and $\delta^{18}\text{O}_{\text{c}}$ for the experiment with 50% meltwater release into the North Atlantic and 50% around Antarctica; **a**, $\delta^{18}\text{O}_{\text{w}}$, and **b**, $\delta^{18}\text{O}_{\text{c}}$. **c, d**, Simulated $\delta^{18}\text{O}_{\text{c}}$ for experiments with 100% meltwater release into the North Atlantic (**c**) and 100% around Antarctica (**d**). Circles give values (or wider ranges in

the case of composite circles) observed in sediment records, as discussed in Supplementary Information. In the vertical profiles, observations have been laterally projected onto the 25° W transect. In the right-hand panels, the *x* axes indicate latitude, with negative values for the Southern Hemisphere.

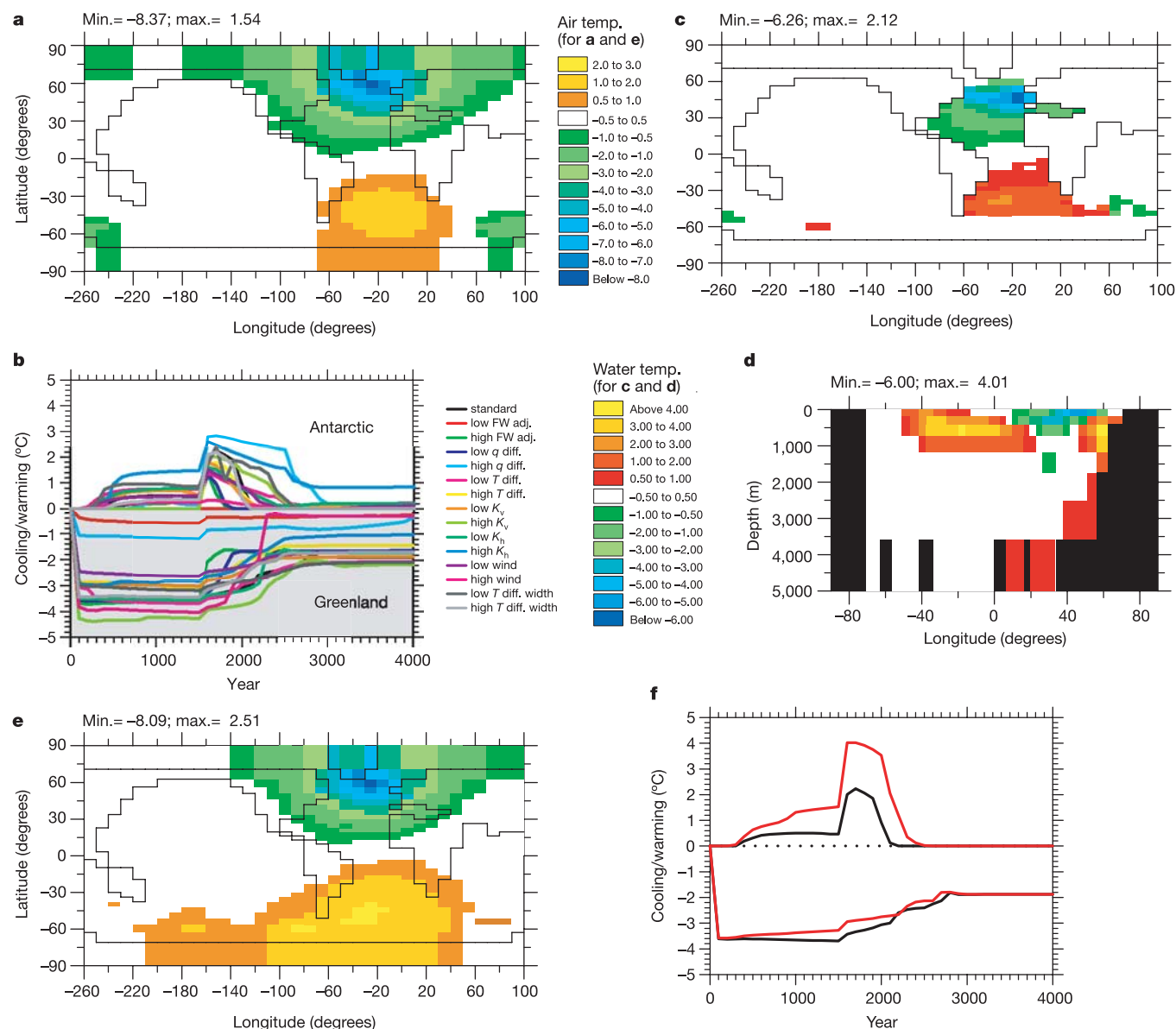


Figure 4 Temperature simulations (at year 1,000) in the 50:50 model, for 0.23 Sv meltwater release. **a**, Air temperature anomaly. **b**, Air temperature development (averages between 20° and 60° W at 70–90° N and 70–90° S) plotted against time in a suite of simulations for the 50:50 model using a wide variety of model parameter settings, illustrating robustness of the result (Methods). FW concerns Atlantic–Pacific exchange of moisture, q is atmospheric moisture diffusivity, T controls diffusive atmospheric heat transport, and K_v and K_h are diffusivity coefficients for diapycnal and isopycnal mixing,

respectively. **c**, Ocean temperature anomaly at 80 m depth. **d**, Ocean temperature anomaly in a section along 25° W longitude. **e**, As **a** but for a scenario with atmospheric $[CO_2]$ increasing linearly by 20 p.p.m.v. over 1,500 yr, and then decreasing linearly to its initial value by year 3000, in a schematic representation of the $[CO_2]$ changes associated with Antarctic temperature changes²⁸. **f**, Plot comparing air temperature changes as in **b**, contrasting the results including (red) and excluding (black) the $[CO_2]$ changes described for **e**.

roughly 500-yr lag between mwp 1A (refs 16, 17) and the Bølling–Allerød interstadial to suggest a comparable causal relationship through changes in the global thermohaline circulation¹⁷.

The inferred roughly 15 m sea level magnitude of the investigated Antarctic ice-volume reductions is towards the high end of current estimates for the glacial ‘excess’ ice volume on Antarctica^{20–24}. It implies that virtually the entire glacial excess might have been involved in millennial-scale variability. □

Methods

The EMIC used here is the three-dimensional ocean/atmosphere/sea ice model C-GOLDSTEIN version 1 (refs 25, 26). The ocean component is a frictional geostrophic model with eddy-induced and isopycnal mixing—appropriate physics to model the global-scale thermohaline circulation properly. Simplified physics and low resolution (36×36 equal-area cells with eight vertical levels) make the model highly efficient. The ocean is coupled to a two-dimensional energy and moisture balance atmosphere and a dynamic

and thermodynamic sea-ice component²⁶.

The model is spun up under simple glacial boundary conditions: atmospheric CO_2 concentration is held constant at 210 p.p.m., and planetary albedo over land poleward of 50° N is increased by 0.18 to mimic ice sheets. To account for the generally low glacial sea level position, the Bering Strait is kept closed in the glacial runs, including all meltwater experiments. The values for 12 key model parameters are the posterior ensemble-mean estimates obtained in a recent tuning exercise, using an ensemble Kalman filter²⁵. In a ‘standard’ spin-up, the model is integrated for 4,000 yr, sufficiently long to equilibrate the global thermohaline circulation (THC). Under glacial boundary conditions, the model simulates air temperatures up to 15°C lower (at high latitudes), and Atlantic sinking of similar strength but shifted 5–10° south compared with a present-day climate simulation.

To investigate post-collapse recovery of the Atlantic overturning (see below), we performed 14 additional spin-up experiments in which 7 of the 12 key parameters (those exerting most control on the THC) were increased or decreased in turn by 2 standard deviations of the ensemble-mean estimates. The Atlantic overturning across these spin-up experiments varies in the range 3–23 Sv (although 12 of the 14 experiments lie in the relatively narrow range 12–20 Sv; see illustrations in Supplementary Information).

From the end of each spin-up, we performed a series of 4,000-yr meltwater release

experiments, in each case adding 0.23 Sv to the local surface freshwater flux over 1,500 yr and 'tagging' the melt water with a $\delta^{18}\text{O}$ value of -40‰ (the background $\delta^{18}\text{O}$ value is assumed zero everywhere). The model offers insight into the likely spatial and depth-dependent distributions of meltwater-induced $\delta^{18}\text{O}$ anomalies, while accounting for the impacts of ocean dynamics on the result. Sensitivity tests have been undertaken with $\delta^{18}\text{O}$ values of -30‰ and -50‰ . The main experiments considered a series of different meltwater release scenarios: 100% in the North Atlantic zone $50\text{--}70^\circ\text{N}$ (additional experiments considered partial or complete release into the Gulf of Mexico); a 50:50 split between the North Atlantic and around Antarctica; 100% around Antarctica; and 20:80, 30:70 and 40:60 splits between the North Atlantic and Antarctica. For diagnostic purposes we averaged $\delta^{18}\text{O}$ over four major basins (Atlantic, Southern Ocean, Pacific, Indian) in four layers (upper, 0–411 m; intermediate, 411–1,158 m; deep, 1,158–2,520 m; bottom, 2,520–5,000 m).

The model has a simplified atmosphere, no carbon cycle (that is, no CO_2 and CH_4 feedbacks), no interactive ice sheets and no insolation changes or solar variability. The scenarios presented are part of a suite of experiments in which the proportions between North Atlantic and Antarctic meltwater combinations were varied in 10% increments. The scenarios comprise constant and simultaneous meltwater fluxes into the North Atlantic and (equally distributed) around Antarctica. As part of the sensitivity analyses, some scenarios have also been run with differently phased northern and southern inputs. In total, nearly 150 4,000-yr simulations have been performed to ensure that the results discussed in this paper are robust with regard to all feasible combinations for the key model parameters.

In the EMIC, the Atlantic overturning is very sensitive to the location of imposed meltwater releases, similarly to previous studies^{17,29,30}. In our extreme case with 100% meltwater release into the North Atlantic, the Atlantic overturning circulation collapses rapidly (Fig. 2), preserving a strong local $\delta^{18}\text{O}$ signal in surface layers during, and shortly after, the period of freshwater release (Fig. 3a). The same collapse (and corresponding isotopic signal) is obtained with every combination of key model parameters considered here (Supplementary Fig. 4). Similar tests for robustness have been made for the 50:50, 40:60 and 30:70 scenarios. These experiments confirm the robustness of our main results, namely the collapse of Atlantic overturning under a sustained local freshwater flux anomaly of about 0.1 Sv, in agreement with other studies^{29,30}.

Received 3 March; accepted 16 July 2004; doi:10.1038/nature02859.

- Langway, C. C. Jr, Oeschger, H. & Dansgaard, W. (eds) *Greenland Ice Core: Geophysics, Geochemistry, and the Environment* (Geophys. Monogr. 33, Am. Geophys. Union, Washington DC, 1985).
- Groote, P. M., Stuiver, M., White, J. W. C., Johnsen, S. & Jouzel, J. Comparison of oxygen isotope records from the GISP2 and GRIP Greenland ice cores. *Nature* **366**, 552–554 (1993).
- Dansgaard, W. *et al.* Evidence for general instability of past climate from a 250 kyr ice core. *Nature* **364**, 218–219 (1993).
- Rahmstorf, S. Timing of abrupt climate change: a precise clock. *Geophys. Res. Lett.* **30** doi:10.1029/2003GL017115 (2003).
- Broecker, W. S. Abrupt climate change: causal constraints provided by the paleoclimate record. *Earth Sci. Rev.* **51**, 137–154 (2000).
- Voelker, A. H. *et al.* Global distribution of centennial-scale records for marine isotope stage (MIS) 3: a database. *Quat. Sci. Rev.* **21**, 1185–1212 (2002).
- Leuschner, D. C. & Sirocko, F. The low-latitude monsoon climate during Dansgaard–Oeschger cycles and Heinrich events. *Quat. Sci. Rev.* **19**, 243–254 (2000).
- Rohling, E. J., Mayewski, P. A. & Challenor, P. On the timing and mechanism of millennial-scale climate variability during the last glacial cycle. *Clim. Dyn.* **20**, 257–267 (2003).
- Pahne, K., Zahn, R., Elderfield, H. & Schulz, M. 340,000-year centennial-scale marine record of Southern Hemisphere climatic oscillation. *Science* **301**, 948–952 (2003).
- Blunier, T. *et al.* Asynchrony of Antarctic and Greenland climate change during the last glacial period. *Nature* **394**, 739–743 (1998).
- Hemming, S. R. Heinrich events: Massive detritus layers of the North Atlantic and their global climate impact. *Rev. Geophys.* **42**, RG1005, doi: 10.1029/2003RG 000128 (2004).
- Shackleton, N. J., Hall, M. A. & Vincent, E. Phase relationships between millennial-scale events 64,000–24,000 years ago. *Paleoceanography* **15**, 565–569 (2000).
- Siddall, M. *et al.* Sea-level fluctuations during the last glacial cycle. *Nature* **423**, 853–858 (2003).
- Cutler, K. B. *et al.* Rapid sea-level fall and deep-ocean temperature change since the last interglacial period. *Earth Planet. Sci. Lett.* **206**, 253–271 (2003).
- Kanfoush, S. L. *et al.* Millennial-scale instability of the Antarctic ice sheet during the last glaciation. *Science* **288**, 1815–1818 (2000).
- Clark, P. U., Mitrovica, J. X., Milne, G. A. & Tamisiea Sea-level fingerprinting as a direct test for the source of global meltwater pulse 1A. *Science* **295**, 2438–2441 (2002).
- Weaver, A. J., Saenko, O. A., Clark, P. U. & Mitrovica, J. X. Meltwater pulse 1A from Antarctica as a trigger of the Bolling–Allerød warm interval. *Science* **299**, 1709–1713 (2003).
- Stocker, T. F. & Johnsen, S. J. A minimum thermodynamic model for the bipolar seesaw. *Paleoceanography* **18**, PA1087, doi:10.1029/2003PA000920 (2003).
- Ganopolski, A. & Rahmstorf, S. Rapid changes of glacial climate simulated in a coupled climate model. *Nature* **409**, 153–158 (2001).
- Barrett, P. J. Antarctic palaeoenvironment through Cenozoic times—a review. *Terra Antarctica* **3**, 103–119 (1996).
- Peltier, W. R. Ice Age palaeotopography. *Science* **265**, 195–201 (1994).
- Huybrechts, P. Sea-level changes at the LGM from ice-dynamic reconstructions of the Greenland and Antarctic ice sheets during the glacial cycles. *Quat. Sci. Rev.* **21**, 203–231 (2002).
- Denton, G. H. & Hughes, T. J. Reconstructing the Antarctic Ice Sheet at the Last Glacial Maximum. *Quat. Sci. Rev.* **21**, 193–202 (2002).
- Anderson, J. B., Shipp, S. S., Lowe, A. L., Wellner, J. S. & Mosola, A. B. The Antarctic Ice Sheet during the Last Glacial Maximum and its subsequent retreat history: a review. *Quat. Sci. Rev.* **21**, 49–70 (2002).
- Annan, J. D., Hargreaves, J. C., Edwards, N. R. & Marsh, R. Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. *Ocean Model.* (in the press).
- Edwards, N. R. & Marsh, R. Uncertainties due to transport parameter sensitivity in an efficient 3-D ocean–climate model. *Clim. Dyn.* (submitted) (2004).

- Stenni, B. *et al.* A late-glacial high-resolution site and source temperature record derived from the EPICA Dome C isotope records (East Antarctica). *Earth Planet. Sci. Lett.* **217**, 183–195 (2003).
- Indermühle, A., Monnin, E., Stauffer, B., Stocker, T. F. & Wahlen, M. Atmospheric CO_2 concentration from 60 to 20 kyr BP from the Taylor Dome ice core, Antarctica. *Geophys. Res. Lett.* **27**, 735–738 (1999).
- Rahmstorf, S. Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycles. *Nature* **378**, 145–149 (1995).
- Rahmstorf, S. Rapid climate transitions in a coupled ocean–atmosphere model. *Nature* **372**, 82–85 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Stocker, J. Thomson, A.P. Roberts, W. Broecker and N. Shackleton for suggestions. All authors contributed equally to this work.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.J.R. (e.rohling@soc.soton.ac.uk).

A Middle Jurassic 'sphenosuchian' from China and the origin of the crocodylian skull

James M. Clark¹, Xing Xu^{2*}, Catherine A. Forster³ & Yuan Wang²

¹Department of Biological Sciences, George Washington University, Washington DC 20052, USA

²Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, 142 Xiwai Street, Beijing, 100044, China

³Department of Anatomical Sciences, Stony Brook University, Stony Brook, New York 11794, USA

* Present address: Department of Paleontology, American Museum of Natural History, 79th Street at Central Park West, New York 10024, USA

The skull of living crocodylians is highly solidified and the jaw closing muscles are enlarged¹, allowing for prey capture by prolonged crushing between the jaws. Living species are all semi-aquatic, with sprawling limbs and a broad body that moves mainly from side-to-side²; however, fossils indicate that they evolved from terrestrial forms. The most cursorial of these fossils^{3–6} are small, gracile forms often grouped together as the Sphenosuchia, with fully erect, slender limbs; their relationships, however, are poorly understood^{5,7–10}. A new crocodylomorph from deposits in northwestern China of the poorly known Middle Jurassic epoch possesses a skull with several adaptations typical of living crocodylians. Postcranially it is similar to sphenosuchians but with even greater adaptations for cursoriality in the forelimb. Here we show, through phylogenetic analysis, that it is the closest relative of the large group Crocodyliformes, including living crocodylians. Thus, important features of the modern crocodylian skull evolved during a phase when the postcranial skeleton was evolving towards greater cursoriality, rather than towards their current semi-aquatic habitus.

The Sphenosuchia includes nine monotypic genera comprising the most basal members of the Crocodylomorpha, but its monophyly has been controversial. Sphenosuchians were once⁷ thought to be a paraphyletic assemblage with some taxa closer to the large group Crocodyliformes. Later evidence suggested that Sphenosuchia is a monophyletic group^{5,8–10}, but some analyses found no resolution owing to conflicting data⁶. Members of the Crocodyliformes are diagnosed by a large suite of features¹¹, some of which are related to the solidifying of intracranial joints and an increase in the

size of muscles responsible for producing a powerful bite force. The new taxon possesses several of these features and greatly reduces the morphological gap between sphenosuchians and crocodyliforms.

Sphenosuchians are found on every continent except Antarctica and Australasia and in deposits from the Late Triassic to the Early Jurassic, and possibly the Late Jurassic^{6,12}. These small (<1.5 m total length) animals are among the most gracile of non-flying archosaurs. The hindlimb is known in several sphenosuchian taxa in which its morphology indicates that it moved within a vertical, parasagittal plane^{3,4}; however, the forelimb and the axial skeleton are not as well known or studied. The new specimen, comprising the

front half of a skeleton, was discovered during recent fieldwork that we directed at Wucuiwan, Xinjiang¹³. It is the most complete skeleton of a non-marine crocodylomorph from the Middle Jurassic.

The skeleton possesses four features typical of the large group Crocodylomorpha^{6,9,10}: a broad squamosal overhanging the temporal region; a medially shifted quadrate contacting the lateral surface of the braincase; a ventrally elongate coracoid; and an elongate radiale and ulnare. It includes typical sphenosuchian features^{9,10} such as the thin, ventrally concave squamosal with a curved posterolateral edge, a compact carpus and a posteroventrally curving rod-like ventral process of the coracoid. Although it lacks most features diagnostic of the Crocodyliformes¹¹—a heavily sculptured skull, a shorter rostrum, a flat skull roof, a smaller antorbital fossa and fenestra, and an anteroposteriorly expanded ventral end of the coracoid—it possesses several features previously not found in sphenosuchians.

Our phylogenetic analysis places *Junggarsuchus* as the sister-group of crocodyliforms, identifying five unambiguous synapomorphies (Fig. 1). Of trees that are two steps longer than ours, some have a monophyletic Sphenosuchia including *Junggarsuchus*, and bootstrap support for the *Junggarsuchus*-crocodyliform node is 61%. Unambiguous synapomorphies, identified by our phylogenetic analysis, include the following (character numbers in brackets are from the phylogenetic analysis, see Supplementary Information): (20) exoccipitals meet on the midline above the foramen magnum; (36) the large ventrolateral extension of the exoccipital contacts the quadrate; (37) the jugal is strongly arched dorsally; (42) the occipital portion of the parietal is narrow; and (46) the quadrate is fenestrated. *Dibothosuchus* (7, 22, 41) and *Sphenosuchus* (16, 19) are nested with this group on the basis of the following unambiguous synapomorphies, all from the skull: (7) the prefrontal does not underlie the frontal; (22) the mastoid antrum enters deeply into the prootic; (41) a horizontal shelf in the posterior part of the supra-temporal fenestra; (16) the parietals are fused; and (19) the occipital margin is straight.

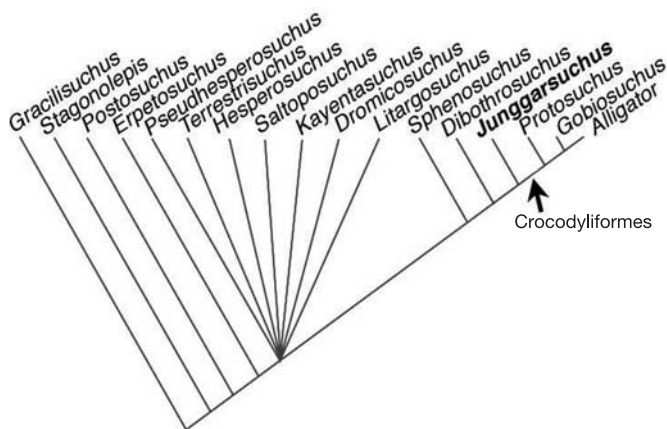


Figure 1 Results of a maximum parsimony phylogenetic analysis. Strict consensus of 81 trees; length = 109, consistency index (CI) = 0.569, retention index (RI) = 0.659.

Junggarsuchus is the closest relative of the Crocodyliformes and the Sphenosuchia are paraphyletic. See Supplementary Information for details of the analysis.

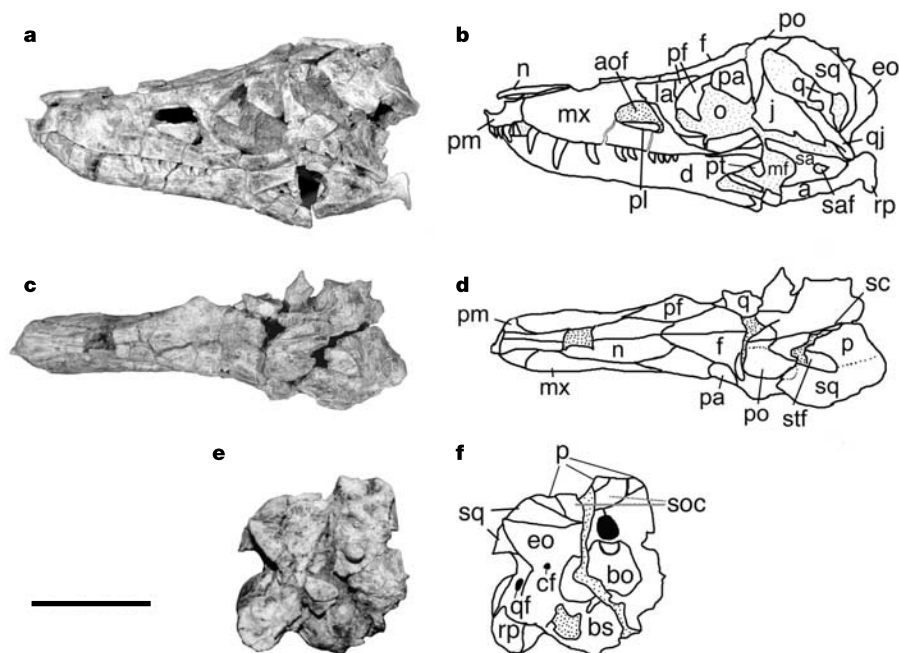


Figure 2 *Junggarsuchus sloani* holotype skull. **a, b**, Left lateral view. **c, d**, Dorsal view. **e, f**, Occipital view. a, angular; aof, antorbital fenestra; bo, basoccipital; bs, basisphenoid; cf, carotid foramen; d, dentary; eo, exoccipital; f, frontal; j, jugal; la, lacrimal; mf, external mandibular fenestra; mx, maxilla; n, nasal; o, orbit; p, parietal; pa, palpebral; pf, prefrontal; pl, palatine; pm, premaxilla; po, postorbital; pt, pterygoid; q, quadrate; qf, quadrate fenestra; qj, quadratojugal; rp, retroarticular process; sa, surangular; saf, surangular foramen; sc, sagittal crest; soc, supraoccipital; sq, squamosal; stf, superior temporal fenestra. Scale bar, 5 cm.

The skull (Fig. 2) is notable for its lack of kinesis and the reduction in the supratemporal fenestrae, as in crocodyliforms. A ventrolateral extension of the otoccipital connects with the distal end of the quadrate, solidifying the temporal region and preventing any movement of the quadrate. Broadened areas for muscle attachment on the arched jugal and surangular indicate an increased size of the *Musculus adductor mandibulae externus pars superficialis*. As in living crocodylians, but not the most basal crocodyliforms or spheosuchians, the lateral surface of the angular has a large area for the insertion of the *M. pterygoideus posterior* extending onto the anterolateral surface of the retroarticular process.

Procoelous vertebral joints, a lack of osteoderms, and trunk vertebrae with short transverse processes and sub-vertical zygapophyses indicate that the body had a greater capacity for vertical movement than in living crocodylians. The lack of osteoderms on the specimen is unlikely to be due to taphonomic loss or ontogenetic immaturity because the rest of the skeleton is preserved undisturbed and the fusion of neurocentral sutures indicates adulthood¹⁴. In living crocodylians the osteoderms are tightly

bound to the axial skeleton through tendinous connections to large epaxial muscles originating on long transverse processes². Together with the nearly horizontal zygapophyseal articulations of the dorsal vertebrae this system greatly restricts vertical mobility and facilitates undulating lateral movements. *Junggarsuchus* vertebrae in the anterior trunk and shoulder regions have small or no transverse processes, and zygapophyses that are oriented less than 30 degrees from the vertical (Fig. 3).

The configuration of the shoulder and wrist and the reduction of the outer fingers indicate that the forelimb was held directly beneath the body. In living crocodylians the shoulder joint involves a posterolaterally facing, notch-shaped glenoid fossa and an elongate, convex humeral head (Fig. 3). This 'hemi-sellar' type of joint is present in all living amniotes except mammals¹⁵. The glenoid fossa of *Junggarsuchus* instead involves a ventrally or posteroventrally facing concave scapular component superficially similar to that of mammals and a smaller, posteriorly facing much shallower coracoid component. The humerus of *Junggarsuchus* is unlike that of any other crocodylomorph in that it has a well-developed hemispherical head projecting perpendicular to the shaft, with a convex proximal articulating surface. The convex head must have articulated with the concave scapular glenoid fossa and have been buttressed anteriorly by the coracoid, so that the shaft of the humerus was vertical or posteroventrally inclined.

The wrist indicates that movement of the hand was in line with the rest of the forelimb and not splayed out as in living crocodylians. In living crocodylians the wrist forms a complex joint in which the broadly convex distal end of the ulna articulates with both proximal carpal bones, which rotate against the ulna when walking, spreading the hand. In *Junggarsuchus* the distal end of the ulna is squared off, and the contact for the radiale on the medial side of the ulna is perpendicular to the rounded distal articulation surface with the ulnare, limiting rotation against the ulna. In living crocodylians there is only one distal carpal and it lies laterally so that the carpal-metacarpal joint is oblique¹⁶; in *Junggarsuchus* the two distal carpals are flattened, equally thick and form a straight joint perpendicular to the long axes of the radiale and ulnare.

The first digit is absent and the fifth digit was relatively small, flexing towards the other toes rather than facing the ground, making the hand functionally tridactyl. In living crocodylians the five fingers are usually spread out lateral to the forearm during locomotion and all face the ground. The first three digits are robust and the other two are smaller. In *Junggarsuchus* the four metacarpals are like those of spheosuchians in that they are compressed together proximally and not spread out^{5,9}, suggesting the manus was digitigrade. Metacarpal V is about half the thickness of metacarpals II–IV and shortened proximally so that it does not reach the carpus. The distal articulating surface of metacarpal V is oriented such that the phalanx flexes towards the other digits rather than posteroventrally, and the metacarpal joint surface is not ginglymoid like those of II–IV. Metacarpal IV is similar in size to metacarpal II, so that the metacarpals of the three weight-bearing digits (II–IV) are symmetrical in size.

The orientation of the glenoid fossa, the convex head on the humerus, the lack of rotation around the distal end of the ulna, the compact metacarpals and reduced outer digits all indicate that the limb moved in a vertical, parasagittal plane. Reduction of the outer digits is typical of highly cursorial mammals¹⁷.

The paraphyly of spheosuchians implies that the postcranial features shared among these taxa are ancestral for crocodyliforms rather than specializations of 'Spheosuchia'. The posteroventrally oriented, concave glenoid fossa on the scapula, the elongate process on the coracoid, the compact carpus and the extremely slender limbs were gained in spheosuchians and then lost in crocodyliforms. The 'Spheosuchia' was thus a phase in the evolution of crocodylomorphs during which the group as a whole became highly adapted for terrestriality. The procoelous vertebrae, lack of osteo-

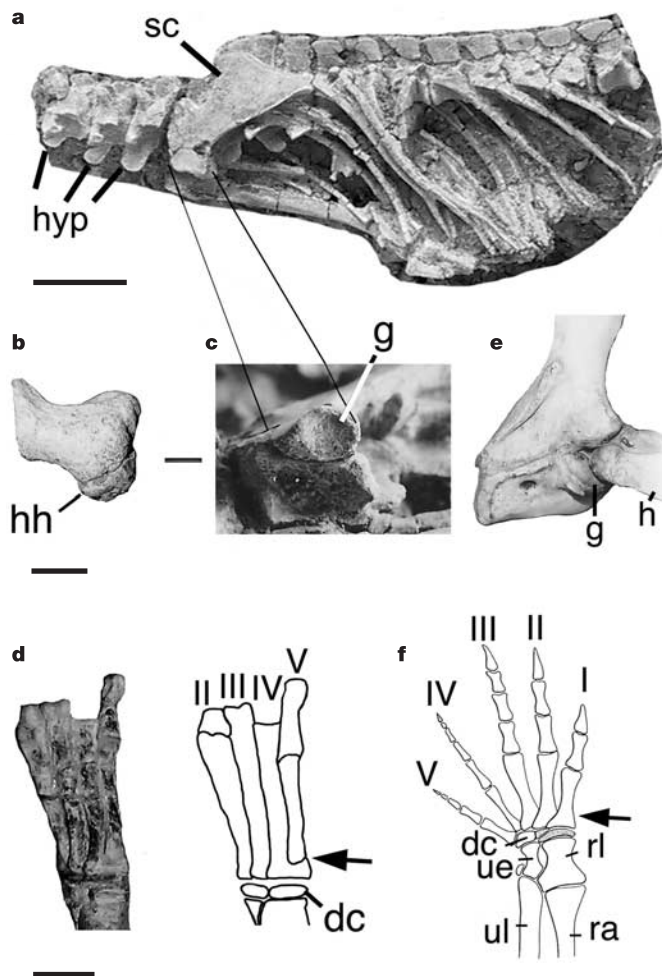


Figure 3 Postcranial skeleton of *Junggarsuchus sloani*. **a**, Skeleton in left lateral view. **b**, Left humerus with hemispherical head in proximal view. **c**, Posteroventrally facing glenoid fossa of left scapula in distal view. **d**, Left carpus and manus in posteroventral view (arrow points to reduced fifth metacarpal). **e**, **f**, For comparison, the living *Alligator mississippiensis* left humerus and shoulder in lateral view (**e**) showing the hemi-sellar glenoid fossa, and left hand in dorsal view (**f**) showing the asymmetry of the digits and carpals, the rounded ulna–carpal contact (arrow) and the single distal carpal (stippled area is cartilaginous). dc, distal carpal; g, glenoid fossa; h, humerus; hh, head of humerus; hyp, hypapophyses; ra, radius; rl, radiale; sc, scapula; ue, ulnare; ul, ulna; I–V, digits of hand. **f** was modified from ref. 16. Scale bars, 4 cm (**a**) and 1 cm (**b–d**).

derms, inturned humeral head and reduced digits of *Junggarsuchus* indicate that the closest known relative of crocodyliforms was also the most highly adapted cursor. The consolidation of the crocodylian skull thus began well before crocodylians entered the water. □

Methods

Order Crocodylomorpha Hay, 1930 *sensu* Walker, 1970³

Junggarsuchus sloani gen. et sp. nov.

Etymology. For the Junggar Basin in northern Xinjiang, *souchous* (Greek) meaning crocodile and for C. Sloan, who discovered the holotype.

Holotype. IVPP V14010, (Institute of Vertebrate Paleontology and Paleoanthropology, Beijing), the articulated anterior half of a skeleton including a nearly complete skull (Figs 2 and 3).

Locality and age. Lower part of the Shishugou Formation¹⁸ at Wucuiwan, Altay Prefecture, Xinjiang. The lower Shishugou (also known as the Wucuiwan Formation) is considered late Middle Jurassic (Bathonian–Callovian)^{19,20}.

Diagnosis. Small (body length ~1 m) non-crocodyliform crocodylomorph with longitudinal concavity on ventrolateral surface of dorsally arched jugal, broadened dorsal edge of surangular, shallow fossa on distal edge of paroccipital process and part of squamosal, enlarged anterior maxillary teeth, well developed surangular foramen, retroarticular process lacking medial process and with broad dorsolateral and posteroventral flanges, shallow procoely in all preserved vertebrae, anteroposteriorly and distally elongate hypapophyses on posterior 4 cervical and anterior 4 dorsal vertebrae, broadened posterior border and sinusoidal dorsal edge on scapula, anteriorly directed humeral head and reduced deltopectoral crest, reduced metacarpal V not contacting carpus, no manus digit I, and no osteoderms.

Phylogenetic analysis

Phylogenetic relationships were analysed using a matrix of 55 characters from the skull and postcranial skeleton distributed among 17 taxa (see Supplementary Information). The taxa included four outgroups and three representatives of the Crocodyliformes. The matrix is derived from earlier studies¹⁰ with 16 additional characters and the addition of *Junggarsuchus*, *Erpetosuchus*²¹ and *Gobiosuchus*²². The data were analysed in PAUP²³.

Received 20 May; accepted 2 July 2004; doi:10.1038/nature02802.

1. Iordansky, N. N. in *Biology of the Reptilia* (eds Gans, C. & Parsons, T. S.) Vol. 4 201–262 (Academic, New York, 1973).
2. Salisbury, S. W. & Frey, E. in *Crocodylian Biology and Evolution* (eds Grigg, G. C., Seebacher, F. & Franklin, C. E.) 85–134 (Surrey Beatty and Sons, Chipping Norton, Australia, 2001).
3. Walker, A. D. A revision of the Jurassic reptile *Hallopus victor* (Marsh), with remarks on the classification of crocodiles. *Phil. Trans. R. Soc. Lond. B* **257**, 323–372 (1970).
4. Parrish, J. M. The origin of crocodylian locomotion. *Paleobiol.* **13**, 396–414 (1987).
5. Sereno, P. C. & Wild, R. *Procompsognathus*: theropod, 'thecodont' or both? *J. Vert. Paleontol.* **12**, 435–458 (1992).
6. Clark, J. M. & Sues, H.-D. Two new species of basal crocodylomorphs and the status of the *Sphenosuchia*. *Zool. J. Linn. Soc.* **136**, 77–96 (2002).
7. Benton, M. J. & Clark, J. M. in *The Phylogeny and Classification of the Tetrapods* (ed. Benton, M. J.) Vol. 1 295–338 (Clarendon, London, 1988).
8. Wu, X.-C. & Chatterjee, S. *Dibothrosuchus elaphros*, a crocodylomorph from the Lower Jurassic of China and the phylogeny of the Sphenosuchia. *J. Vert. Paleontol.* **13**, 58–89 (1993).
9. Clark, J. M., Sues, H.-D. & Berman, D. S. A new specimen of *Hesperosuchus agilis* from the Upper Triassic of New Mexico and the interrelationships of basal crocodylomorph archosaurs. *J. Vert. Paleontol.* **20**, 683–704 (2001).
10. Sues, H.-D., Olsen, P. E., Carter, J. G. & Scott, D. M. A new crocodylomorph archosaur from the Upper Triassic of North Carolina. *J. Vert. Paleontol.* **23**, 329–343 (2003).
11. Clark, J. M. in *In the Shadow of the Dinosaurs: Early Mesozoic Tetrapods* (eds Fraser, N. & Sues, H.-D.) 84–97 (Cambridge Univ. Press, New York, 1994).
12. Ague, J. J., Carpenter, K. & Ostrom, J. H. Solution to the *Hallopus* enigma? *Am. J. Sci.* **295**, 1–17 (1995).
13. Clark, J. M. & Xu, X. Dinosaurs from Xinjiang, China (<http://www.gwu.edu/~clade/faculty/clark/china.html>) (2001–2003).
14. Brochu, C. A. Closure of neurocentral sutures during crocodylian ontogeny: implications for maturity assessment in fossil archosaurs. *J. Vert. Paleontol.* **16**, 49–62 (1996).
15. Jenkins, F. A. Jr The evolution of the avian shoulder joint. *Am. J. Sci.* **A** **293**, 253–267 (1993).
16. Müller, G. & Alberch, P. Ontogeny of the limb skeleton in *Alligator mississippiensis*: developmental invariance and change in the evolution of archosaur limbs. *J. Morphol.* **203**, 151–164 (1990).
17. Hildebr, M. & Goslow, G. *Analysis of Vertebrate Structure* (Wiley, New York, 2001).
18. Eberth, D. A. *et al.* Sequence stratigraphy, paleoclimate patterns and vertebrate fossil distributions in Jurassic–Cretaceous strata of the Junggar Basin, Xinjiang Autonomous Region, PR China. *Can. J. Earth Sci.* **38**, 1627–1644 (2001).
19. Chen, P. J. in *The Continental Jurassic* (ed. Morales, M.) 395–412 (Museum of Northern Arizona Bulletin 60, Flagstaff, 1996).
20. Lucas, S. G. *Chinese Fossil Vertebrates* (Columbia Univ. Press, New York, 2001).
21. Benton, M. J. & Walker, A. D. *Erpetosuchus*, a crocodile-like basal archosaur from the Late Triassic of Elgin, Scotland. *Zool. J. Linn. Soc.* **136**, 25–47 (2002).
22. Osmólska, H., Hua, S. & Buffetaut, E. *Gobiosuchus kielanae* (Protosuchia) from the Late Cretaceous of Mongolia: anatomy and relationships. *Acta Palaeontol. Pol.* **42**, 257–289 (1997).
23. Swofford, D. L. PAUP* Beta 10 Software (Sinauer Associates, Sunderland, Massachusetts, 2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements Field work was supported by the National Geographic Society, the National Natural Science Foundation of China, the Jurassic Foundation, the Hilmar Sallee bequest, George Washington University and the Chinese Academy of Sciences. Study of the specimen was supported by the National Science Foundation Division of Earth Sciences. We thank D. Ma and W. Chen of the Changji Autonomous Prefecture and M. Zhu and X. Zhao of the IVPP for facilitating our work.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.C. (jclark@gwu.edu).

Cooperation and competition in pathogenic bacteria

Ashleigh S. Griffin¹, Stuart A. West¹ & Angus Buckling²

¹Institute of Cell, Animal & Population Biology, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3JT, UK

²Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

Explaining altruistic cooperation is one of the greatest challenges for evolutionary biology^{1–3}. One solution to this problem is if costly cooperative behaviours are directed towards relatives^{4,5}. This idea of kin selection has been hugely influential and applied widely from microorganisms to vertebrates^{2–10}. However, a problem arises if there is local competition for resources, because this leads to competition between relatives, reducing selection for cooperation^{3,11–14}. Here we use an experimental evolution approach to test the effect of the scale of competition, and how it interacts with relatedness. The cooperative trait that we examine is the production of siderophores, iron-scavenging agents, in the pathogenic bacterium *Pseudomonas aeruginosa*^{15–17}. As expected, our results show that higher levels of cooperative siderophore production evolve in the higher relatedness treatments. However, our results also show that more local competition selects for lower levels of siderophore production and that there is a significant interaction between relatedness and the scale of competition, with relatedness having less effect when the scale of competition is more local. More generally, the scale of competition is likely to be of particular importance for the evolution of cooperation in microorganisms, and also the virulence of pathogenic microorganisms, because cooperative traits such as siderophore production have an important role in determining virulence^{6,9,17–19}.

Explaining altruistic cooperation is fundamental to understanding the main evolutionary transitions from single-celled organisms to complex animal societies¹. The problem is why should an individual carry out an altruistic behaviour that is costly to perform, but benefits another individual or the local group? Hamilton's^{4,5} kin selection theory provides an explanation for altruism between relatives: by helping a close relative reproduce, an individual is still passing on its genes to the next generation indirectly. This is encapsulated by Hamilton's rule^{3,4}, which states that an altruistic behaviour is favoured whenever $rb - c > 0$, where r is the genetic relatedness between the actor and the beneficiary, b is the benefit of receiving the altruistic behaviour and c is the cost of performing the behaviour. The theory of kin selection is well accepted, and variation in relatedness has been applied to explain variation in the level of altruistic cooperative behaviours in organisms ranging from microorganisms to vertebrates^{2–10}.

However, selection for altruism depends also upon the scale of competition (population demography or structure)^{3,11}. There is a

large theoretical literature on this topic, initiated by the debate on whether limited dispersal (population viscosity) favours the evolution of altruism^{2,3,11–14,20}. Although limited dispersal leads to a higher relatedness, r , between interacting individuals, which favours altruism, it can also lead to more local competition between relatives. This reduces the advantage of altruism because the increased fitness of the relative who receives the altruism is increasingly paid for by other relatives¹³. Frank³ showed that this effect could be incorporated into Hamilton's rule, by allowing the scale at which competition occurs to vary from global, with no competition between relatives, to local, with potentially extreme competition between relatives. Specifically, he used an extra parameter, a , which is the proportion of competition that is local, and can vary from $a = 0$ (global competition) to $a = 1$ (local competition). This leads to an extended version of Hamilton's rule³, which predicts that the

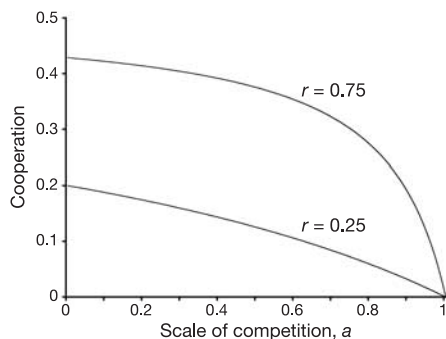


Figure 1 Scale of competition and kin selection theory. We have plotted the effect of the scale of competition on selection for an altruistic trait, from the incorporation of Frank's³ scale of competition parameter, a , into a classic tragedy of the commons formulation^{3,17}. This allows a simple and general graphical representation of the theoretical predictions. The y-axis gives the evolutionary stable allocation of resources to a cooperative trait that is costly, but provides a benefit locally^{3,25}. The scale of competition varies from global ($a = 0$) to local ($a = 1$). The different lines represent relatively high ($r = 0.75$) and relatively low ($r = 0.25$) relatedness. Higher levels of cooperation are favoured when relatedness is higher (higher r), and competition is more global (lower a). Furthermore, there is an interaction between scale of competition and relatedness: as competition becomes more local, the influence of relatedness on selection for cooperation is reduced. In the extreme, if $a = 1$, then competition is completely local and so kin selection cannot favour altruism^{3,22}. The same conclusions can be reached using Queller's¹¹ approach of allowing for local competition, or with a model specifically developed for siderophore production¹⁷.

scale of competition has two influences. First, as competition becomes more local (higher a), cooperation is selected against (Fig. 1). Second, there is an interaction between scale of competition and relatedness: as competition becomes more local, the influence of relatedness on selection for cooperation is reduced (Fig. 1). These effects occur because local competition leads to increased competition between neighbours and relatives, decreasing the kin-selected benefit of altruism^{3,11,13,20}.

Unfortunately, this problem is often ignored, across taxa ranging from microparasites^{3,21} to cooperative breeding vertebrates¹², and Hamilton² left its importance as a major unsolved problem. To an extent this is because there is a lack of empirical evidence for how the scale of competition can influence selection for altruistic cooperative behaviours¹². Empirical tests have been hindered because the amount of competition between relatives is usually confounded with relatedness; both increase with reduced dispersal^{13,14,22}. Consequently, the best support for the role of scale of competition has come from other areas of social evolution and cases of extremely local competition, such as the favouring of female-biased sex ratios when males compete locally for mates^{2,3,11,12,22}. An ideal experiment to address this problem would involve the independent manipulation of relatedness, r , and the scale of competition, a .

Here we use a bacterial system to carry out such a study, by experimentally testing the effects of the scale of competition and relatedness on the evolution of altruistic cooperation. Bacteria offer exceptional opportunities for addressing this problem because the effects of population demography on relatedness and the scale of competition can be disentangled, and because evolutionary responses can be followed over a reasonable timescale^{9,23,24}. The system that we use is siderophore production in the opportunistic pathogen *P. aeruginosa*. Iron is a major limiting factor for bacterial growth because most iron in the environment is in the insoluble Fe(III) form, and, in the context of bacterial parasites, is actively withheld by hosts^{15–17}. Siderophores are agents produced by bacteria in response to iron deficiency, that are released into the environment to scavenge insoluble and host-bound iron, making it available for bacterial metabolism^{15–17}. Siderophore production is an altruistic cooperative trait that is costly for the individual but provides a local (group) benefit because other individuals can take up the siderophore–iron complex (see Methods)^{15–17}. It is therefore termed a 'whole group' cooperative trait²⁵. Importantly, *P. aeruginosa* mutants that do not produce siderophores have been seen to evolve both *in vitro* and in the lungs of cystic fibrosis patients^{17,26}. Furthermore, the relationship between siderophore

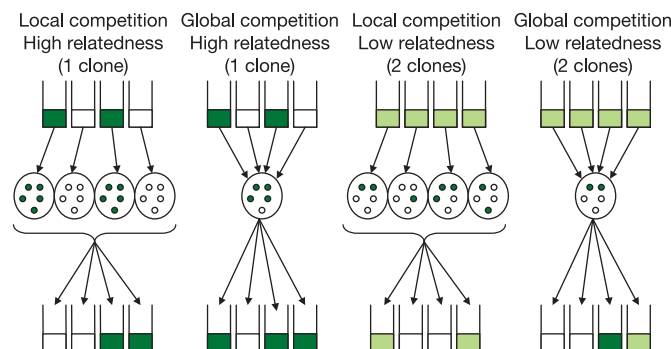


Figure 2 Experimental design. We varied relatedness between interacting individuals by initiating each subpopulation with either a single bacterial clone (relatively high r) or two bacterial clones (relatively low r). We varied the scale of competition by either mixing the cultures from all of the subpopulations in a treatment before plating, and then transferring random colonies from this single plate to initiate new subpopulations (relatively global

competition, lower a), or by allowing every subpopulation in a treatment to provide equal numbers of colonies to the next generation (relatively local competition, higher a). We use dark green to symbolize the siderophore-producing cooperator, white to symbolize a cheater that does not produce siderophores, and light green for a mixture of cooperators and cheaters.

production and the growth rates of parasitic bacteria suggests that siderophore production plays a key part in determining parasite virulence, the damage to a host resulting from parasite infection^{15–17,27}.

We performed an experiment in which we independently manipulated relatedness and the scale of competition, and determined the consequences for the evolution of cooperation. We used a classic two factorial analysis of variance (ANOVA) design, and set up four treatments in which relatedness was either relatively high or low and the scale of competition was either relatively global or local (Fig. 2). Each treatment was replicated four times, giving a total of 16 (4×4) independent replicates. Each replicate contained one population divided into 12 subpopulations. The evolution of altruistic siderophore production was monitored by starting with a 50:50 overall ratio of cooperators and cheats, and monitoring how this changed over time. The cooperator is a wild-type strain that produces the primary siderophore, pyoverdine, and the cheater is a mutant strain that does not. These can be readily distinguished because pyoverdine is green, and so colours the wild-type cooperator colonies, whereas pyoverdine-minus cheater colonies are white.

To achieve relatively high relatedness (higher r) between individuals, we initiated each subpopulation with a single bacterial clone: in the first generation half the subpopulations in a treatment were initiated with the cooperator strain that produces pyoverdine, and the other half with the cheater strain that does not (Fig. 2). In contrast, we imposed relatively low relatedness (lower r) by initiating each subpopulation with two bacterial clones: in the first generation all of the subpopulations were initiated with a 50:50 mix of cooperators and cheaters. Note that relatedness is measured with respect to the 'altruist allele' (pyoverdine production), which is the relevant measure³. We imposed relatively global competition (lower a) by mixing the cultures from all of the subpopulations in a treatment before plating, and then transferring random colonies from this single plate to initiate new subpopulations (Fig. 2). This procedure allows productivity in a tube to determine the genetic contributions to subsequent generations, increasing the relative importance of global (between subpopulation) competition. In contrast, we imposed relatively local competition (higher a) by allowing every subpopulation in a treatment to provide equal numbers of colonies to the next generation. This removes the advantage of being in a more productive tube, and hence increases

the importance of local (within subpopulation) competition.

As expected from Hamilton's rule⁴, cooperative siderophore producers were favoured in the higher relatedness relative to the lower relatedness treatments ($F_{(1,13)} = 73.6$, $P < 0.0001$; Fig. 3). This occurs in our experiment because the low relatedness treatments provide a relative advantage to cheating, by allowing cheaters to exploit siderophores produced by cooperators whenever a subpopulation contains both types. However, as explicitly predicted by Frank's³ extension of kin selection theory (Fig. 1), the success of cooperators was also dependent on the scale of competition, and its interaction with relatedness. First, cooperators had an advantage under relatively global competition compared with relatively local competition ($F_{(1,13)} = 44.8$, $P < 0.0001$; Fig. 3). Global competition provides a relative advantage to cooperators, even when relatedness is low, because by chance there will be some subpopulations containing only cooperators, which are more productive and hence provide a greater contribution to the next generation. Second, there was a significant interaction between relatedness and the scale of competition, with relatedness having a relatively weaker effect when the scale of competition was more local ($F_{(1,12)} = 7.4$, $P = 0.019$; Fig. 3). The variation in cooperator success was associated with a significant increase in cooperator frequency (one treatment), a significant increase in cheater frequency (two treatments), and no significant change from starting conditions (one treatment) (Fig. 3). Overall, relatedness, the scale of competition and their interaction explained 94% of the variance in cooperator frequency.

Our results provide a clear experimental demonstration of how the scale of competition influences the evolution of altruistic cooperation, both in isolation and through its interaction with relatedness. The scale of competition could reduce selection for cooperation between relatives across a wide range of organisms¹²; for example, it is likely to be important in some cooperative breeding vertebrates¹², and could help explain the variation across species in the extent to which individuals preferentially help relatives⁸. We suggest that the scale of competition is likely to be of particular importance for the evolution of cooperation in bacteria and other microorganisms. This is because life cycles can involve stages with limited movement and hence relatively local growth and competition, but also relatively long distance dispersal stages^{6,9,28}. Consequently, the overall scale of competition is likely to vary continuously across species, between local and global, depending on dispersal rates. For example, in the context of bacterial pathogens this variation will occur as a result of within-host growth (relatively local) and transmission to new hosts (relatively global)^{3,21}.

Many traits associated with increased growth and virulence in pathogenic bacteria seem to be altruistic and subject to kin selection. Here we have discussed siderophore production, but other examples include biofilms, Shiga toxins and immune suppression^{17–19}. The link between cooperation and virulence could be exploited with new intervention strategies, and also suggests that virulence could evolve as a consequence of intervention²⁹. For example, if the reduced transmission from a control programme led to more local competition (higher a), without significantly altering relatedness, then it would select for less cooperation and hence lower virulence. These consequences could be particularly important in pathogens such as *P. aeruginosa*, in which transmission can occur between infected sufferers of cystic fibrosis³⁰, and evolution can take place on such a short timescale relative to the duration of an infection²⁶. □

Methods

Strains

P. aeruginosa ATC 15692 (PA01) was used as the pyoverdine (siderophore) producer, and strain PA06609, a mutant derived by UV-mutagenesis from a methionine auxotroph (generated by transposon-mutagenesis) of PA01, was used as the pyoverdine-negative cheat²⁷.

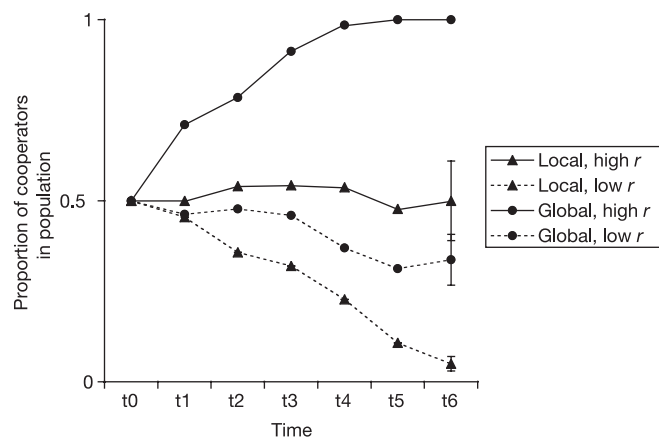


Figure 3 The evolution of cooperation in response to relatedness and the scale of competition. The proportion of cooperating individuals who produce pyoverdine siderophores is plotted against time. The different lines represent relatively high (solid line) and low (dashed line) relatedness. The different symbols represent relatively global (circle) and local (triangle) competition. Each of the four treatments was replicated four times, and standard errors are shown for the final time point. Time is measured as transfers, between which cultures were allowed to grow for 24 h. Cooperation is favoured by higher relatedness and more global competition.

Benefit and cost of siderophores

We determined that siderophore production is an appropriate cooperative trait for our experiment. We measured the growth rates of a wild-type strain that produces the primary siderophore, pyoverdine (cooperator), and a mutant strain that does not (cheater), either when alone or in mixed cultures where the other strain was present, and at a variety of iron availabilities. Both strains were grown overnight in 30 ml glass universals containing 6 ml standard King's medium B (KB) in an orbital shaker (200 r.p.m.), at 37 °C. Cell densities did not differ between strains when grown under these conditions ($n = 12$, 2-sample t -test: $t = 0.37$, $P > 0.2$). Sixteen universals containing Casamino acids medium (CAA; 5 g Casamino acids, 1.18 g $K_2HPO_4 \cdot 3H_2O$, 0.25 g $MgSO_4 \cdot 7H_2O$, per litre) media were inoculated with 10^6 overnight culture cells of either PA01, PA6609 or a 50:50 mixture of both, resulting in a total of 48 cultures. Sodium bicarbonate was added to each tube to create a 20 mM solution, necessary for effective chelator activity²⁷. Four tubes inoculated with the three different bacterial populations (PA01, PA6609 and both) were exposed to three different iron-limitation treatments: 50 $\mu g\ ml^{-1}$, 100 $\mu g\ ml^{-1}$ and 200 $\mu g\ ml^{-1}$ of human apo-transferrin (Sigma), a natural iron-chelator, were added. Human iron chelators (such as transferrin) bind iron, preventing non-siderophore-mediated uptake of iron by bacteria. Siderophores are also strong iron chelators, and so compete with transferrin for iron. The tubes were incubated at 37 °C in a static incubator for 24 h. Cultures were then plated onto KB agar and relative densities were determined.

Siderophore (pyoverdine) production in these strains has the characteristics that make an appropriate cooperative trait for our study. First, there is a cost to individuals that can produce pyoverdine, as shown by monocultures of mutants being able to outcompete wild-type strains in an iron-rich environment ($F_{(1,6)} = 12.61$, $P = 0.01$). Second, when iron is limiting, siderophore production is beneficial, as shown by populations containing wild-type strains growing to a higher density than those where pyoverdine producers are absent ($F_{(1,21)} = 16.13$, $P = 0.0006$). Chelator concentration was also significant ($F_{(1,21)} = 12.45$, $P = 0.002$), but not the interaction ($F_{(1,20)} = 0.69$, $P = 0.42$). Third, individuals who do not produce pyoverdine are able to exploit the pyoverdine produced by the others, as shown by mutants growing to higher densities in the presence of wild type in iron-poor environments ($t = 6.37$, d.f. = 10, $P < 0.005$). Consequently, pyoverdine production is a costly altruistic trait that can potentially be exploited by cheaters.

Selection experiment

We independently manipulated relatedness and the scale of competition, using a classic two factorial ANOVA design. Replicates contained one population divided into 12 subpopulations. Each subpopulation was grown in a tube of CAA medium supplemented with 20 mM $NaHCO_3$ and 100 $\mu g\ ml^{-1}$ human apo-transferrin²⁷. In high relatedness treatments, six tubes were inoculated with 10^6 cells from an overnight culture of either PA01 or PA6609 (1:1 overall ratio of altruist to cheater). Low relatedness treatments initially comprised 12 tubes inoculated with 10^6 cells of a 1:1 mix of both strains. Cultures were then grown for 24 h in a 37 °C static incubator, during which time approximately seven generations take place. Local competition cultures were then individually plated onto KB agar, whereas an equal volume from each culture within a global competition treatment were mixed together before plating. Plates were incubated for 24 h at 37 °C, and after determining relative frequencies of the two strains, 24 or 12 (low and high relatedness treatments, respectively) random colonies were separately inoculated into KB media and grown at 200 r.p.m. and 37 °C, overnight. CAA tubes were then inoculated with a total of 10^6 cells from these overnight cultures: low relatedness tubes were inoculated with one clone, and high relatedness with two. (Colonies were grown up in KB rather than used to inoculate the subsequent transfer directly, to control for the difference in size between altruistic and cheater colonies, and therefore starting densities). This selection procedure was repeated for six transfers, allowing a total of 42 (7×6) bacterial generations under experimental conditions. Every round of selection we counted the frequencies of altruistic and selfish bacteria growing on the agar plates, and the relative proportion inoculated into the next round. We used the proportion of cooperators inoculated into the next generation as the response variable in our analyses, and in Fig. 3. The whole experiment was repeated a further three times.

Analyses

We analysed our data using standard ANOVA methodology, as implemented in the package GLMStat 5.7.5 (Kagi Shareware). For all analyses on the proportion of cheaters (pyoverdine-minus colonies) in the population, the proportion was arcsine square root transformed before analysis, and a normal distribution subsequently confirmed with a Shapiro-Wilkinson test. We present the analyses for the end of the experiment after six transfers.

Received 26 April; accepted 11 June 2004; doi:10.1038/nature02744.

- Maynard Smith, J. & Szathmari, E. *The Major Transitions in Evolution* (W.H. Freeman, Oxford, 1995).
- Hamilton, W. D. *Narrow Roads of Gene Land: I Evolution of Social Behaviour* (W.H. Freeman, Oxford, 1996).
- Frank, S. A. *Foundations of Social Evolution* (Princeton Univ. Press, Princeton, 1998).
- Hamilton, W. D. The evolution of altruistic behaviour. *Am. Nat.* **97**, 354–356 (1963).
- Hamilton, W. D. The genetical evolution of social behaviour, I & II. *J. Theor. Biol.* **7**, 1–52 (1964).
- Crespi, B. J. The evolution of social behavior in microorganisms. *Trends Ecol. Evol.* **16**, 178–183 (2001).
- Keller, L. & Reeve, H. K. in *Encyclopedia of Evolution* (ed. Pagel, M. D.) 595–600 (Oxford Univ. Press, Oxford, 2002).
- Griffin, A. S. & West, S. A. Kin discrimination and the benefit of helping in cooperatively breeding vertebrates. *Science* **302**, 634–636 (2003).
- Velicer, G. J. Social strife in the microbial world. *Trends Microbiol.* **11**, 330–337 (2003).
- Queller, D. C. & Strassmann, J. E. Kin selection and social insects. *Bioscience* **48**, 165–175 (1998).
- Queller, D. C. Genetic relatedness in viscous populations. *Evol. Ecol.* **8**, 70–73 (1994).

- West, S. A., Pen, I. & Griffin, A. S. Cooperation and competition between relatives. *Science* **296**, 72–75 (2002).
- Taylor, P. D. Altruism in viscous populations - an inclusive fitness model. *Evol. Ecol.* **6**, 352–356 (1992).
- Wilson, D. S., Pollock, G. B. & Dugatkin, L. A. Can altruism evolve in purely viscous populations. *Evol. Ecol.* **6**, 331–341 (1992).
- Guerinot, M. L. Microbial iron transport. *Annu. Rev. Microbiol.* **48**, 743–772 (1994).
- Ratledge, C. & Dover, L. G. Iron metabolism in pathogenic bacteria. *Annu. Rev. Microbiol.* **54**, 881–941 (2000).
- West, S. A. & Buckling, A. Cooperation, virulence and siderophore production in bacterial parasites. *Proc. R. Soc. Lond. B* **270**, 37–44 (2003).
- Brown, S. P. Cooperation and conflict in host-manipulating parasites. *Proc. R. Soc. Lond. B* **266**, 1899–1904 (1999).
- Brown, S. P., Hochberg, M. E. & Grenfell, B. T. Does multiple infection select for raised virulence? *Trends Microbiol.* **10**, 401–405 (2002).
- Grafen, A. in *Behavioural Ecology: An Evolutionary Approach* (eds Krebs, J. R. & Davies, N. B.) 62–84 (Blackwell Scientific Publications, Oxford, 1984).
- Frank, S. A. Models of parasite virulence. *Q. Rev. Biol.* **71**, 37–78 (1996).
- West, S. A., Murray, M. G., Machado, C. A., Griffin, A. S. & Herre, E. A. Testing Hamilton's rule with competition between relatives. *Nature* **409**, 510–513 (2001).
- Velicer, G. J. & Yu, Y. N. Evolution of novel cooperative swarming in the bacterium *Myxococcus xanthus*. *Nature* **425**, 75–78 (2003).
- Rainey, P. B. & Rainey, K. Evolution of cooperation and conflict in experimental bacterial populations. *Nature* **425**, 72–74 (2003).
- Pepper, J. W. Relatedness in trait group models of social evolution. *J. Theor. Biol.* **206**, 355–368 (2000).
- De Vos, D. et al. Study of pyoverdine type and production by *Pseudomonas aeruginosa* isolated from cystic fibrosis patients: prevalence of type II pyoverdine isolates and accumulation of pyoverdine-negative mutants. *Arch. Microbiol.* **175**, 384–388 (2001).
- Meyer, J. M., Neely, A., Stintzi, A., Georges, C. & Holder, I. A. Pyoverdine is essential for virulence of *Pseudomonas aeruginosa*. *Infect. Immun.* **64**, 518–523 (1996).
- Strassmann, J. E., Zhu, Y. & Queller, D. C. Altruism and social cheating in the social amoeba *Dictyostelium discoideum*. *Nature* **408**, 965–967 (2000).
- Gandon, S., Mackinnon, M. J., Nee, S. & Read, A. F. Imperfect vaccines and the evolution of pathogen virulence. *Nature* **414**, 751–756 (2002).
- Hoogkamp-Korstanje, J. A. A., Meis, J. F. G. M., Kissing, J., van der Laag, J. & Melchers, W. J. G. Risk of cross-colonization and infection by *Pseudomonas aeruginosa* in a holiday camp for cystic fibrosis patients. *J. Clin. Microbiol.* **33**, 572–575 (1995).

Acknowledgements We thank J.-M. Meyer for supplying strains; A. Gardner and D. Shuker for discussion and comments; A. Duncan and A. Graham for laboratory assistance; staff at the Bega Public Library, NSW, Australia, for internet access; BBSRC, NERC and Royal Society for funding.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.S.G. (a.griffin@ed.ac.uk).

High rates of N₂ fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean

Joseph P. Montoya¹, Carolyn M. Holl¹, Jonathan P. Zehr², Andrew Hansen³, Tracy A. Villareal⁴ & Douglas G. Capone⁵

¹School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

²Department of Ocean Sciences and Institute of Marine Science, University of California, Santa Cruz, California 95064, USA

³School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, Hawaii 96822, USA

⁴Marine Science Institute, The University of Texas at Austin, Port Aransas, Texas 78373, USA

⁵Department of Biological Sciences and The Wrigley Institute for Environmental Studies, University of Southern California, Los Angeles, California 90089, USA

The availability of nitrogen is important in regulating biological productivity in marine environments. Deepwater nitrate has long been considered the major source of new nitrogen supporting primary production in oligotrophic regions of the open ocean, but recent studies have showed that biological N₂ fixation has a critical role in supporting oceanic new production^{1–7}. Large colonial cyanobacteria in the genus *Trichodesmium* and the

heterocystous endosymbiont *Richelia* have traditionally been considered the dominant marine N_2 fixers, but unicellular diazotrophic cyanobacteria and bacterioplankton have recently been found in the picoplankton and nanoplankton community of the North Pacific central gyre, and a variety of molecular and isotopic evidence suggests that these unicells could make a major contribution to the oceanic N budget⁸. Here we report rates of N_2 fixation by these small, previously overlooked diazotrophs that, although spatially variable, can equal or exceed the rate of N_2 fixation reported for larger, more obvious organisms. Direct measurements of $^{15}N_2$ fixation by small diazotrophs in various parts of the Pacific Ocean, including the waters off Hawaii where the unicellular diazotrophs were first characterized, show that N_2 fixation by unicellular diazotrophs can support a significant fraction of total new production in oligotrophic waters.

The recent discovery of N_2 -fixing coccoid cyanobacteria and bacteria in the North Pacific subtropical gyre⁸ revealed a potentially important source of new nitrogen to support oceanic production, but the small size and low numeric abundance of these organisms⁸ pose significant challenges for quantifying their N_2 fixation activity. We applied a high-sensitivity tracer technique⁹ to measure the rate of N_2 fixation by the small size fraction (less than $10\mu m$) of plankton from an extensive region of subtropical and tropical waters in the Pacific. In contrast to two recent studies that documented measurable but low rates of N_2 fixation by small

diazotrophs at or near the Hawaii Ocean Timeseries station ALOHA north of Oahu^{10,11}, our data provide the first spatially extensive direct measurements of N_2 fixation by small diazotrophs in oligotrophic waters and show that these recently discovered organisms can make a major contribution to oceanic N_2 fixation.

Experiments performed across broad expanses of the eastern North Pacific and waters north of Australia show that the rate of $^{15}N_2$ fixation by the unicellular diazotroph size fraction is measurable, and at times very high (Fig. 1). Our most intensive experiments were performed with water collected at 25 m depth at ALOHA, or from the surface outside Kaneohe bay, Oahu. In all these experiments, we found N_2 fixation by small diazotrophs at rates of 0.01 – $0.15\text{ nmol N l}^{-1}\text{ h}^{-1}$ (Fig. 2a), with the highest rates in August 2001 and February 2002. By integrating our rates through the water column (see Methods), we calculated areal rates of N_2 fixation between 11 and $103\mu\text{mol N m}^{-2}\text{ d}^{-1}$ with a mean of $66 \pm 19\mu\text{mol N m}^{-2}\text{ d}^{-1}$ ($N = 7$). Although we conservatively assumed only 12 h of daily N_2 fixation activity, our data do not show any consistent temporal phasing of N_2 fixation activity, and N_2 fixation seems to occur during the day as well as at night. This result contrasts sharply with the obligate diel pattern in N_2 fixation activity by *Trichodesmium*^{12,13}, and the apparent dominance of night-time acetylene reduction in highly concentrated samples taken near station ALOHA¹¹.

The volumetric rates of N_2 fixation that we measured for ALOHA

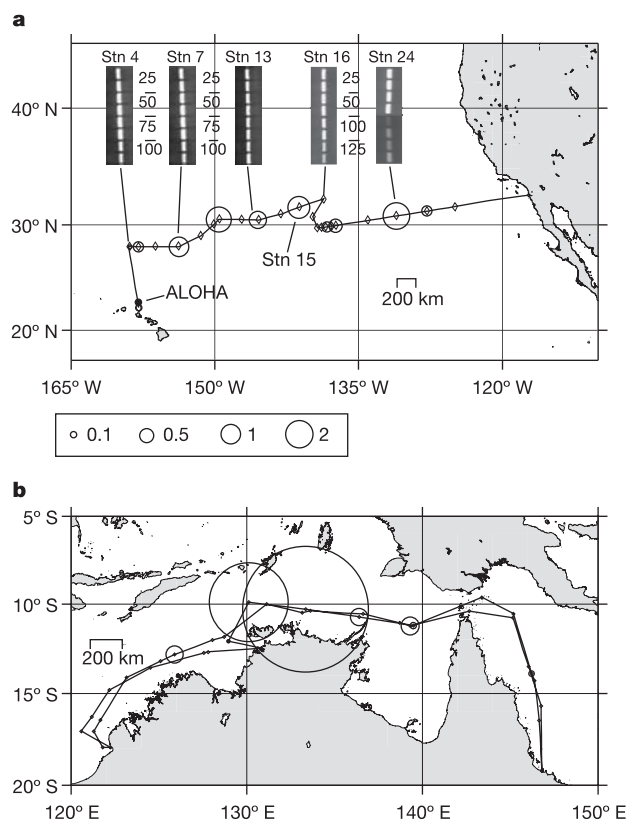


Figure 1 Volumetric rate of N_2 fixation by unicellular organisms in the North Pacific Ocean and near Australia. The area of each circle is proportional to the N_2 fixation rate as shown (in $\text{nmol N l}^{-1}\text{ h}^{-1}$). Stations where no rate measurements were made are marked with diamonds. **a**, N_2 fixation rates measured at ALOHA and during cruise Cook-25 (Jun – Jul 2002). Insets show occurrence of *nifH* (bands) as a function of depth in metres at selected stations. At all depths, clone libraries constructed by amplification with degenerate *nifH* primers were dominated by group A cyanobacterial sequences. **b**, Rates measured during cruise EW9912 (October to November 1999). Charts prepared using the Generic Mapping Tools^{20,21}.

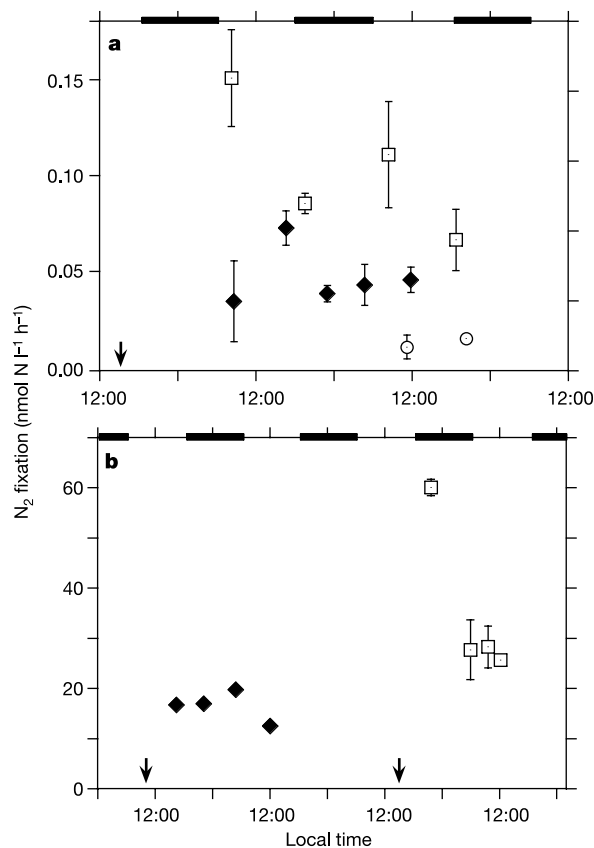


Figure 2 Time course of N_2 fixation in incubations under simulated *in situ* conditions. Each symbol represents the integrated rate of N_2 fixation between the start (arrow) and the endpoint of an incubation. **a**, Experiments performed in July 2000 (circles) and August 2001 (squares) with water from 25 m at ALOHA and an experiment performed in February 2002 (diamonds) with water collected just outside Kaneohe bay. Results are means $\pm$ s.d. for three replicates. **b**, Experiments at two stations in the Arafura Sea using water from the subsurface pigment maximum (50–60 m). Results are means $\pm$ s.d. for two replicates.

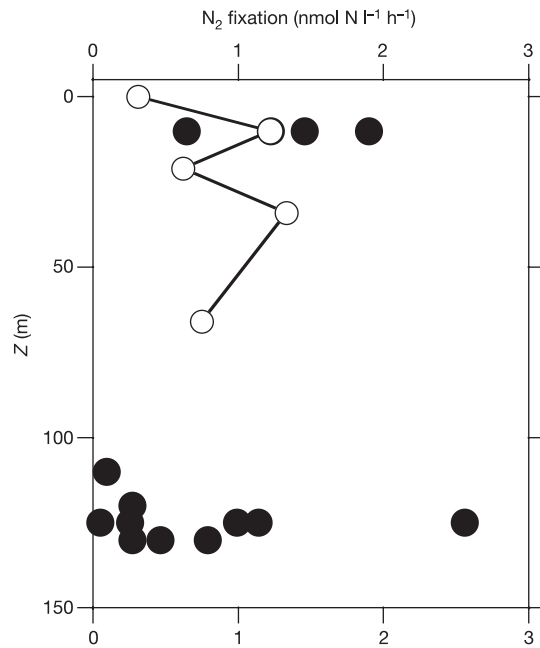


Figure 3 Summary of rate measurements as a function of depth at stations along a transect in the eastern subtropical Pacific from Hawaii to San Diego (cruise Cook-25). At station 15 (31° 32.21' N, 141° 15.66' W), N₂ fixation was measured in water samples from five different depths through the mixed layer (open circles). At other stations, N₂ fixation was measured in water collected either from the subsurface pigment maximum or from about 10 m depth (filled circles).

are substantially higher than those previously reported for the same waters^{10,11}. This contrast might reflect differences in experimental approach and sample handling as well as real temporal variation in N₂ fixation activity in this region. Nitrogenase gene sequences of two distinct groups of unicellular cyanobacteria have previously been detected in the Pacific Ocean at station ALOHA⁸. One group, designated group A, has yet to be fully characterized, whereas the group B sequence type is closely related to cultivated strains of marine *Synechocystis* (*Crocospaera*). Group A diazotrophic cyanobacteria sequence types were recovered from messenger RNA in all experiments with water from station ALOHA. Direct cell counts revealed *Synechocystis*-like cells (that is, potential diazotrophs 3–7 μm in diameter, compared with smaller synechococoids and prochlorococoids) at concentrations of 5–10 cells ml^{-1} in July 2000 and up to 200 cells ml^{-1} in our August 2001 experiment. These differences in *Synechocystis*-like cell abundance are large enough to account for the range of variation that we observed in N₂ fixation rate at ALOHA.

We also performed a series of experiments and diazotroph diversity measurements along a cruise track through the eastern North Pacific from Hawaii to California (Fig. 1a), along which we measured N₂ fixation in water from the mixed layer and subsurface pigment maximum to assess the spatial distribution of diazotrophy across a significant expanse of subtropical waters. On this transect, the maximal pigment concentrations typically occurred near the base of the mixed layer at about 120 m depth, although several stations in a bloom of large diatoms (*Rhizosolenia* and *Hemiaulus*) had shallow pigment maxima at about 10 m depth. At all 11 stations where we performed rate measurements, we found measurable N₂ fixation activity in the pigment maximum, ranging between 0.047 and 1.85 $\text{nmol N l}^{-1} \text{h}^{-1}$, with an overall mean of $0.72 \pm 0.20 \text{ nmol N l}^{-1} \text{h}^{-1}$ ($N = 11$). Most of the rates measured on this

Table 1 Summary of N₂ fixation estimates for oligotrophic waters

Location/domain	Dates	Mean areal rate ($\mu\text{mol N m}^{-2} \text{d}^{-1}$)	S.e.m.	N	Notes
Direct rate measurements					
Unicellular diazotrophs					
ALOHA ¹⁰	2000–2001	24	6	3	¹⁵ N ₂ uptake integrated over 100 m
ALOHA region ¹¹	Autumn 2002	2.2		1	AR in pre-concentrated (130–160 $\times$) water samples
ALOHA (this study)	2000–2001	66	19	7	¹⁵ N ₂ uptake; 77% of the rate at 25 m integrated over 100 m; see text for additional details
Kaneohe bay (this study)	2000–2002	24	6	12	¹⁵ N ₂ uptake integrated over 25 m
Eastern North Pacific gyre (this study)	June–July 2002	520	160	10	¹⁵ N ₂ uptake integrated through the mixed layer and excluding one high outlier (3,830 $\mu\text{mol N m}^{-2} \text{d}^{-1}$ at station 24)
Timor–Arafura–Coral seas (this study)	Nov. 1999	126	47	7	¹⁵ N ₂ uptake integrated through the mixed layer and excluding stations 26 and 27
Arafura Sea (this study)	Nov. 1999	3,955		2	¹⁵ N ₂ uptake in the pigment maximum, stations 26 and 27
Trichodesmium					
Southwestern N Atlantic ²² 0–24° N, 45–66° W	Nov. 1964	41	7.6	19	¹⁵ N ₂ uptake
	May 1965	108	24	17	
Caribbean ²³ , 12–22° N		161	20	12	AR
BATS ²⁴	1995–1997	41		14	¹⁵ N ₂ uptake
N Atlantic	1995–2001	254	40	155	AR (D.G.C., J. A. Burns, J.P.M., A. F. Michaels, A. Subramaniam and E. J. Carpenter, unpublished data)
N Pacific ²⁵ , 21° N, 159° W	1972	134		2	AR
East China Sea ²⁶ , 10–25° N		126	49	32	AR
HOT/ALOHA ⁵	1990–1992	84	43	8	AR
Arabian Sea ²⁷ , 7–10° N	May 1995	35	7.4	9	AR
Arabian Sea ²⁷ , 10° N bloom	May 1995	99	25	5	AR, upper 0.5 m
Richelia/Hemiaulus					
SW N Atlantic ²⁸ , 7–27° N	Oct. 1996	3,110	1,315	14	AR
Geochemical budgets and mass balances					
N Atlantic ⁴		500–2,500			Based on N* distributions
N Atlantic ²		197			Based on N* distributions
Atlantic ¹⁴ , 40° N to 40° S		48			Based on dissolved inorganic C inventory
Pacific ¹⁴ , 40° N to 40° S		137			Based on dissolved inorganic C inventory
Pacific ²⁹		97			Total rate (59 Tg N yr ⁻¹) based on N* distributions; Areal rate calculated for region between 40° N and 40° S.
HOT/ALOHA ⁵		93	38		Based on N:P stoichiometry
HOT/ALOHA ¹⁰		135	15	11	Based on ¹⁵ N mass balance

HOT, Hawaii Ocean Timeseries. BATS, Bermuda Atlantic Timeseries station. N* is a quantitative measure of N:P ratio anomalies relative to oceanic averages. Rate estimates are based on direct measurements as well as geochemical budgets. Acetylene reduction (AR) estimates are based on a 3:1 reduction ratio for conversion to N₂ fixation rates.

transect (Figs 1a and 3) were substantially higher than at ALOHA (Table 1). Although this cruise occurred in a different year from our ALOHA experiments, the rate we measured at the station nearest ALOHA was quite similar to the mean of the rates measured in our focused experiments with ALOHA water (Fig. 1a). Thus, N_2 fixation rates by unicellular diazotrophs might be much higher elsewhere in the basin, and N_2 fixation rates at ALOHA might not be representative of the entire subtropical gyre.

Our rate measurements from this cruise generally focused on a single depth at each station, although we successfully amplified *nifH* DNA sequences throughout the mixed layer of all stations assayed for diazotroph diversity (Fig. 1a). As at ALOHA, our clone libraries from this cruise track were dominated by group A cyanobacterial sequences, whereas bacterial sequences composed only about 1% of the community. The areal rate of N_2 fixation ranged between 70 and 2,800 $\mu\text{mol N m}^{-2} \text{d}^{-1}$. Excluding the one extreme value, the overall mean for the entire transect was $520 \pm 160 \mu\text{mol N m}^{-2} \text{d}^{-1}$ ($N = 10$). In comparison, estimates of the average rate of injection of deepwater NO_3^- into the mixed layer of the oligotrophic North Pacific span a range of about 150–1100 $\mu\text{mol N m}^{-2} \text{d}^{-1}$ (refs 10, 14, 15) in the central gyre to 3600 $\mu\text{mol N m}^{-2} \text{d}^{-1}$ in the equatorial upwelling region¹⁶. This implies that N_2 fixation by small plankton can support a substantial fraction of total new production in the North Pacific and might be comparable to deepwater NO_3^- as a source of new nitrogen to the euphotic zone. The very highest rates of N_2 fixation that we measured could potentially supply new nitrogen to the upper water column at a rate comparable to that of the equatorial upwelling system, although without any accompanying PO_4^{3-} to support the balanced growth of plankton.

Experiments performed on a cruise north of Australia complement our North Pacific data (Fig. 2b). In this shallow shelf system we found measurable N_2 fixation by small diazotrophs at all stations where we conducted rate measurements (Fig. 1b). In particular, extremely high N_2 fixation rates were found in the pigment maximum (50–60 m) in the Arafura Sea. In two experiments performed on successive days, we measured average rates of N_2 fixation of 20–30 $\text{nmol N l}^{-1} \text{h}^{-1}$ and as high as 62 $\text{nmol N l}^{-1} \text{h}^{-1}$. The latter value is the highest rate we have ever measured in a natural sample and occurred in a 7-h incubation extending from mid-afternoon into the evening (14:00 to 21:00 local) of the second day of this study (Fig. 2b). As at ALOHA, the rate of N_2 fixation was independent of time of day (Fig. 2b) and the rate of N_2 fixation in these two Arafura Sea experiments could turn over the standing stock of nitrogen in the pigment maximum on a timescale of 1–4 days.

We calculated a mean areal rate of N_2 fixation in the mixed layer of 126 $\mu\text{mol N m}^{-2} \text{d}^{-1}$ for locations other than our two high-rate stations in the Arafura Sea (see Methods). This is roughly a quarter of the mean rate that we measured for our North Pacific transect. In contrast, at our two high-activity stations in the Arafura Sea, we calculated an areal rate of N_2 fixation in the pigment maximum proper of almost 4 $\text{mmol N m}^{-2} \text{d}^{-1}$ (Fig. 1b, Table 1).

Unlike *Trichodesmium*, which resides in the shallower portions of the upper euphotic zone¹⁷, unicellular diazotrophs seem more uniformly distributed through the water column (Fig. 1a), and our experiments show clear evidence for significant N_2 fixation by unicells in diverse parts of the Pacific Ocean. Our rates are substantially higher than reported previously for unicellular diazotrophs (Table 1), a contrast that might reflect both spatial/temporal variability and differences in experimental detail^{10,11}. In addition, the errors likely to arise in $^{15}\text{N}_2$ tracer incubations (for example, loss of tracer from the incubation vessel and leakage of fixed nitrogen from cells during filtration) are asymmetric in effect and tend to produce underestimates of the true rate⁹. In sum, our results indicate a major flux of new nitrogen entering oceanic systems in addition to that due to large diazotrophs alone. Moreover, *Trichodesmium* has very few known grazers⁶, so the pathways by

which recently fixed nitrogen enters higher trophic levels might also be distinct for these two groups of marine diazotrophs. Our observations support a fundamental change in our estimates of the rate and mechanisms of nutrient supply to the oligotrophic ocean. Our conservative estimates (for example, assuming only 12 h of activity per day) of the mean N_2 fixation rate by small diazotrophs in the North Pacific, extrapolated over the region with mean annual surface temperature of more than 25 °C (about $29 \times 10^{12} \text{ m}^2$), could support new production of roughly 0.4 Pg C yr^{-1} , which is about 10% of recent estimates of total oceanic new production¹⁴. N_2 fixation by unicellular diazotrophs might be the largest flux of new nitrogen into the upper water column of the oligotrophic Pacific, and perhaps other such waters. □

Methods

Rate measurements

We collected water samples for N_2 fixation assays at the Hawaii Ocean Timeseries station ALOHA, in the waters north of Australia (cruise EW-9912) and along a transect from Hawaii to San Diego (cruise Cook-25). All water samples were passed through 100- μm Nitex mesh before incubation to exclude large organisms, including *Trichodesmium* colonies and large diatoms such as *Hemiaulus* that might contain endosymbiotic N_2 fixers. Measurements of N_2 fixation rate were performed in 4-l polycarbonate bottles equipped with silicone septum caps to which trace additions of $^{15}\text{N}_2$ (99 atom%; Cambridge Isotope) were made with a gas-tight syringe.

For experiments in Hawaii, water was collected either at 25 m depth at ALOHA during routine Hawaii Ocean Timeseries cruises with the use of a CTD-rosette sampling system or at the surface just outside Kaneohe bay, Oahu. Once collected, water samples were transported in 50-l polycarbonate bottles to the Hawaii Institute of Marine Biology on Coconut Island, Kaneohe bay. Once there, the sample was transferred to 4-l incubation bottles for rate measurements. All experimental incubations were run in triplicate and terminated in a time series extending 36–48 h.

During cruises EW9912 (October to November 1997) and Cook-25 (June to July 2002), water samples were collected from the mixed layer and pigment maximum with the use of a CTD-rosette system and were incubated on deck for 24–36 h.

Incubations were terminated by gentle pressure filtration through a precombusted GF/F filter after passage through a 10- μm prefilter to remove larger phytoplankton from the sample. Sample filters were dried at 60 °C, then stored over desiccant until analysed by continuous-flow isotope ratio mass spectrometry with a Micromass Optima mass spectrometer interfaced to a CE Elantech NA2500 elemental analyser.

Areal rate calculations

At most stations we made N_2 fixation measurements at one or several depths and took different approaches to integrating these rates through the water column to estimate the areal rate of N_2 fixation. In experiments carried out at ALOHA¹⁰, the rate of N_2 fixation decreased with depth such that the average rate in the upper 100 m of the water column was about 77% of the rate measured at 25 m. We used this scaling factor (77%) to estimate the areal rate of N_2 fixation implied by our rate measurements at 25 m depth at ALOHA.

We took a more conservative approach in estimating areal fluxes by using measurements from our cruises in the North Pacific and Australian waters. In both cases, our data generally focused on a single depth at each station, typically in the subsurface pigment maximum at 50–100 m. A profile of rate measurements made at station 15 near the middle of the North Pacific transect showed rates of 0.31–1.3 $\text{nmol N l}^{-1} \text{h}^{-1}$ with a mean of 0.85 ± 0.21 ($N = 5$) and little obvious vertical structure (Fig. 3). These rates agree well with the range of values measured in the pigment maximum elsewhere on the transect (Fig. 3), indicating that our measurements in the pigment maximum should be representative of the mixed layer as a whole. While recognizing that vertically resolved rate measurements at all stations would be preferable, we conservatively estimated the areal rate of N_2 fixation on these cruises by assuming that N_2 fixation was uniform through the mixed layer rather than highest at the surface. We then integrated our volumetric rates through the mixed layer and assumed, again conservatively, that N_2 fixation occurs for only 12 h d^{-1} . For the two Arafura Sea stations with extremely high rates of N_2 fixation in the pigment maximum, we performed our rate integration through the pigment maximum alone (10 m), rather than through the entire mixed layer.

Where confidence intervals are given, results are means $\pm$ s.e.m.

Diazotroph diversity

Water samples were collected for molecular characterization of the diazotroph community at ALOHA and at selected stations on the Cook-25 cruise. One-litre seawater samples from each depth were filtered on 0.2 μm pore-size Supor filters (Pall Gelman Inc., Ann Arbor, Michigan). The filters were placed in 2-ml centrifuge tubes containing 500 μl of 1 $\times$ Tris-EDTA buffer (10 mM Tris-HCl at pH 7.4 and 1 mM EDTA at pH 8.0). Centrifuge tubes were immediately frozen in liquid nitrogen and transported to the shore-based laboratory for analyses.

Nucleic acids were extracted with the protocol described by Tillett and Neilan¹⁸. DNA concentrations in the extracts were quantified by PicoGreen DNA quantification (Molecular Probes, Eugene, Oregon) with a spectral fluorimeter in accordance with the manufacturer's specifications. To assess the diversity of *nifH*-containing plankton at

station ALOHA, samples were amplified by polymerase chain reaction (PCR) with a nested set of degenerate *nifH* primers^{19,20}. The resulting PCR products were cloned and sequenced to characterize the *nifH* sequence diversity. Detailed descriptions of the methods used for the PCR amplification, cloning and sequencing are found in ref. 19.

Received 18 May; accepted 8 July 2004; doi:10.1038/nature02824.

1. Carpenter, E. J. & Romans, K. Major role of the cyanobacterium *Trichodesmium* in nutrient cycling in the North Atlantic Ocean. *Science* **254**, 1356–1358 (1991).
2. Gruber, N. & Sarmiento, J. L. Global patterns of marine nitrogen fixation and denitrification. *Glob. Biogeochem. Cycles* **11**, 235–266 (1997).
3. Lipschultz, F. & Owens, N. J. P. An assessment of nitrogen fixation as a source of nitrogen to the North Atlantic Ocean. *Biogeochemistry* **35**, 261–274 (1996).
4. Michaels, A. F. *et al.* Inputs, losses and transformations of nitrogen and phosphorus in the pelagic North Atlantic Ocean. *Biogeochemistry* **35**, 181–226 (1996).
5. Karl, D. *et al.* The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature* **388**, 533–538 (1997).
6. Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B. & Carpenter, E. J. *Trichodesmium*, a globally significant marine cyanobacterium. *Science* **276**, 1221–1229 (1997).
7. Montoya, J. P., Carpenter, E. J. & Capone, D. G. Nitrogen-fixation and nitrogen isotope abundances in zooplankton of the oligotrophic North Atlantic. *Limnol. Oceanogr.* **47**, 1617–1628 (2002).
8. Zehr, J. P. *et al.* Unicellular cyanobacteria fix N₂ in the subtropical North Pacific Ocean. *Nature* **412**, 635–638 (2001).
9. Montoya, J. P., Voss, M., Kaehler, P. & Capone, D. G. A simple, high precision, high sensitivity tracer assay for dinitrogen fixation. *Appl. Environ. Microbiol.* **62**, 986–993 (1996).
10. Dore, J. E., Brum, J. R., Tupas, L. & Karl, D. M. Seasonal and interannual variability in sources of nitrogen supporting export in the oligotrophic subtropical North Pacific Ocean. *Limnol. Oceanogr.* **47**, 1595–1607 (2002).
11. Falcon, L. I., Carpenter, E. J., Cipriano, F., Bergman, B. & Capone, D. G. N₂ fixation by unicellular bacterioplankton from the Atlantic and Pacific Oceans: Phylogeny and in situ rates. *Appl. Environ. Microbiol.* **70**, 765–770 (2004).
12. Capone, D. G., O'Neil, J. M., Zehr, J. & Carpenter, E. J. Basis for diel variation in nitrogenase activity in the marine planktonic cyanobacterium *Trichodesmium thiebautii*. *Appl. Environ. Microbiol.* **56**, 3532–3536 (1990).
13. Chen, Y.-B., Zehr, J. P. & Mellon, M. T. Growth and nitrogen fixation of the diazotrophic filamentous nonheterocystous cyanobacterium *Trichodesmium* sp. IMS101 in defined media: evidence for a circadian rhythm. *J. Phycol.* **32**, 916–923 (1996).
14. Lee, K., Karl, D. M., Wanninkhof, R. H. & Zhang, J. Z. Global estimates of net carbon production in the nitrate-depleted tropical and subtropical oceans. *Geophys. Res. Lett.* **29** (2002) doi: 10.1029/2001GL014198.
15. Allen, C. B., Kanda, J. & Laws, E. A. New production and photosynthetic rates within and outside a cyclonic mesoscale eddy in the North Pacific subtropical gyre. *Deep-Sea Res.* **143**, 917–936 (1996).
16. Aufdenkampe, A. K. *et al.* Biogeochemical controls on new production in the tropical Pacific. *Deep-Sea Res.* **49**, 2619–2648 (2002).
17. Carpenter, E. J., Subramaniam, A. & Capone, D. G. Biomass and productivity of the cyanobacterium, *Trichodesmium* spp. in the tropical North Atlantic Ocean. *Deep-Sea Res.* **51**, 173–203 (2004).
18. Tillett, D. & Neilan, B. A. Xanthogenate nucleic acid isolation from cultured and environmental cyanobacteria. *J. Phycol.* **36**, 251–258 (2004).
19. Zehr, J. P. & Turner, P. J. in *Methods in Marine Microbiology* (ed. Paul, J. H.) 271–286 (Academic, San Diego, 2001).
20. Zehr, J. P. & McReynolds, L. A. Use of degenerate oligonucleotides for amplification of the *nifH* gene from the marine cyanobacterium *Trichodesmium thiebautii*. *Appl. Environ. Microbiol.* **55**, 2522–2526 (1989).
21. Wessel, P. & Smith, W. H. F. New, improved version of Generic Mapping Tools released. *Eos* **79**, 579 (1998).
22. Goering, J. J., Dugdale, R. C. & Menzel, D. W. Estimate of in situ rates of nitrogen uptake by *Trichodesmium* sp. in the Tropical Atlantic Ocean. *Limnol. Oceanogr.* **11**, 614–620 (1966).
23. Carpenter, E. J. & Price, C. C. Nitrogen fixation, distribution and production of *Oscillatoria* (*Trichodesmium*) spp. in the western Sargasso and Caribbean Seas. *Limnol. Oceanogr.* **22**, 60–72 (1977).
24. Orcutt, K. M. *et al.* A seasonal study of the significance of N₂ fixation by *Trichodesmium* spp. at the Bermuda Atlantic Time-series Study (BATS) site. *Deep-Sea Res.* **48**, 1583–1608 (2001).
25. Gundersen, K. R. *et al.* Structure and biological dynamics of the oligotrophic ocean photic zone off the Hawaiian Islands. *Pacif. Sci.* **30**, 45–68 (1976).
26. Saino, T. *Biological Nitrogen Fixation in the Ocean with Emphasis on the Nitrogen Fixing Blue-Green Alga, Trichodesmium, and its Significance in the Nitrogen Cycle in the Low Latitude Sea Areas* (Univ. Tokyo, Tokyo, 1977).
27. Capone, D. G. *et al.* An extensive bloom of the N₂-fixing cyanobacterium, *Trichodesmium erythraeum*, in the central Arabian Sea. *Mar. Ecol. Prog. Ser.* **172**, 281–292 (1998).
28. Carpenter, E. J. *et al.* Extensive bloom of a N₂-fixing diatom/cyanobacterial association in the tropical Atlantic Ocean. *Mar. Ecol. Prog. Ser.* **185**, 273–283 (1999).
29. Deutsch, C. *et al.* Denitrification and N₂ fixation in the Pacific Ocean. *Glob. Biogeochem. Cycles* **15**, 483–506 (2001).

Acknowledgements We thank D. Karl and the entire HOT team for support in Hawaii; A. Gibson for collecting samples for molecular characterization on cruise Cook-25; C. Payne, K. Rathbun, S. Patel, K. Ghanouni, P. Davoodi and G. Stewart for their assistance in the laboratory analyses; and the captains and crews of the RV *Ewing* and RV *Melville* for their assistance at sea. This research was supported by grants from the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.P.M. (j.montoya@biology.gatech.edu).

Control of phyllotaxy by the cytokinin-inducible response regulator homologue *ABPHYL1*

Anna Giulini*, Jing Wang & David Jackson

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA

* Present address: Department of Plant Production, University of Milan, Via Celoria 2, 20133 Milan, Italy

Phyllotaxy describes the geometric pattern of leaves and flowers, and has intrigued botanists and mathematicians for centuries^{1,2}. How these patterns are initiated is poorly understood, and this is partly due to the paucity of mutants³. Signalling by the plant hormone auxin appears to determine the site of leaf initiation; however, this observation does not explain how distinct patterns of phyllotaxy are initiated⁴. *abphyl1* (*abph1*) mutants of maize initiate leaves in a decussate pattern (that is, paired at 180°), in contrast to the alternating or distichous phyllotaxy observed in wild-type maize and other grasses⁵. Here we show that *ABPH1* is homologous to two-component response regulators and is induced by the plant hormone cytokinin. *ABPH1* is expressed in the embryonic shoot apical meristem, and its spatial expression pattern changes rapidly with cytokinin treatment. We propose that *ABPH1* controls phyllotactic patterning by negatively regulating the cytokinin-induced expansion of the shoot meristem, thereby limiting the space available for primordium initiation at the apex.

Named after their ability to promote proliferation of cultured cells, the cytokinins have diverse roles in shoot morphogenesis^{6,7}. Cytokinin signals are perceived by trans-membrane histidine kinase receptors, which regulate gene expression through transcription factors known as response regulators (reviewed in refs 8, 9). Although the molecular details of this pathway have been described in detail, the precise biological functions of the response regulators are unknown, as loss-of-function mutants do not have striking phenotypes^{10–13}. Here we show that the phyllotaxy mutant *abph1* is caused by a loss-of-function mutation in a cytokinin-inducible response regulator.

abph1 mutants can have opposite leaf (decussate) phyllotaxy throughout their life cycle, and also initiate ears in opposite pairs (Fig. 1a, b). The altered phyllotaxy originates at the shoot apical meristem (SAM) by the initiation of leaf primordia in opposite pairs (Fig. 1c, d)^{5,14}. Our reference *abph1* allele, *abph1-0*, arose spontaneously in breeding stocks⁵. We screened for putative tagged alleles in progeny from crosses of mutant plants onto *Mutator* (*Mu*) or *Suppressor-mutator* transposon lines, and recovered four novel alleles. Co-segregation analysis using one of the *Mu* alleles, *abph1-mum181*, identified a 4-kilobase (kb) *Mu7* hybridizing band in 50 mutant individuals but not in any wild-type siblings or progenitors (not shown). We cloned this band from a sub-genomic library, and amplified the flanking sequences by polymerase chain reaction (PCR). The products contained an entire open reading frame arranged in five predicted exons, with the *Mu7* transposon 3 bp downstream of the predicted start codon (Fig. 1e).

We used this candidate gene as a probe to characterize additional *abph1* alleles. Southern blotting revealed that the three other transposon alleles were deleted for the cloned flanking region (Supplementary Fig. S1b); *Mu* transposons are known to give rise to deletions¹⁵. We were unable to amplify the intact coding region from the *abph1-0* reference allele (supplementary data, Fig. S1d) or to detect transcript from this allele on northern blots (not shown) or by PCR with reverse transcriptase (RT-PCR) (Fig. 2a). We obtained

an additional allele from F. Salamini, and detected a single-nucleotide substitution at the 5' junction of the first intron, which led to mis-splicing or non-splicing of this intron (Supplementary Fig. S1e). In summary, we found molecular lesions in six independent *abph1* alleles, confirming that the cloned gene corresponds to *ABPH1*. Each of the alleles is likely to be a null, as we failed to detect normal *ABPH1* transcript in any of them.

The predicted coding sequence of *ABPH1* was identical to *Zea mays* response regulator 3 (*ZmRR3*)¹⁶, which is related to a

cytokinin-inducible response regulator, although expression of that gene was not inducible by cytokinin in leaves. As our first description of *ABPH1* pre-dates this description, we continue to use the name *ABPH1* (ref. 5). The predicted protein product is 15 kDa and is a member of the 'A' class of response regulators that contains a receiver domain but no DNA-binding output domain. An alignment with CheY, a bacterial response regulator, is shown in Fig. 1f, with invariant Asp and Lys residues marked¹⁷. Phylogenetic analysis failed to detect obvious orthology relationships between members

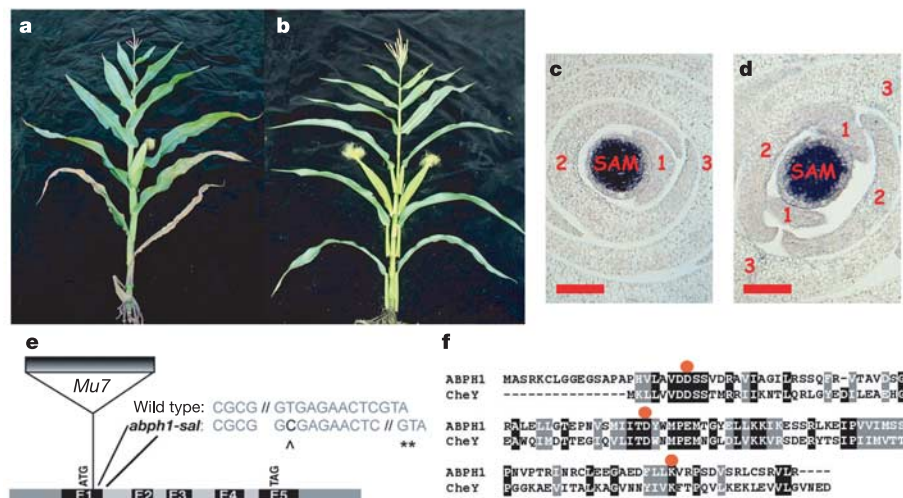


Figure 1 *abph1* phenotypes and gene isolation. **a, b**, Adult normal (**a**) and *abph1* mutant plant (**b**) showing decussate phyllotaxy and two prominent ears in the mutant; some leaves were removed for clarity. **c, d**, The decussate phyllotaxy initiates at the SAM, as seen in transverse section of normal (**c**) and *abph1* mutant (**d**) apices, hybridized *in situ* with *knotted1* probe. Leaf primordia are numbered (with the youngest as 1). Scale bar,

100 μ m. **e**, Map of the *ABPH1* locus showing exons (E1–E5), start ATG and stop TAG codons, the position of the *Mu7* transposon in the *abph1-181* allele, and the sequence of the exon1–intron1 junction in the wild type and *abph1-sal* allele. Asterisks indicate the new intron splice donor site in *abph1-sal*. **f**, Alignment of *ABPH1* and CheY protein sequences, with invariant Asp and Lys residues marked.

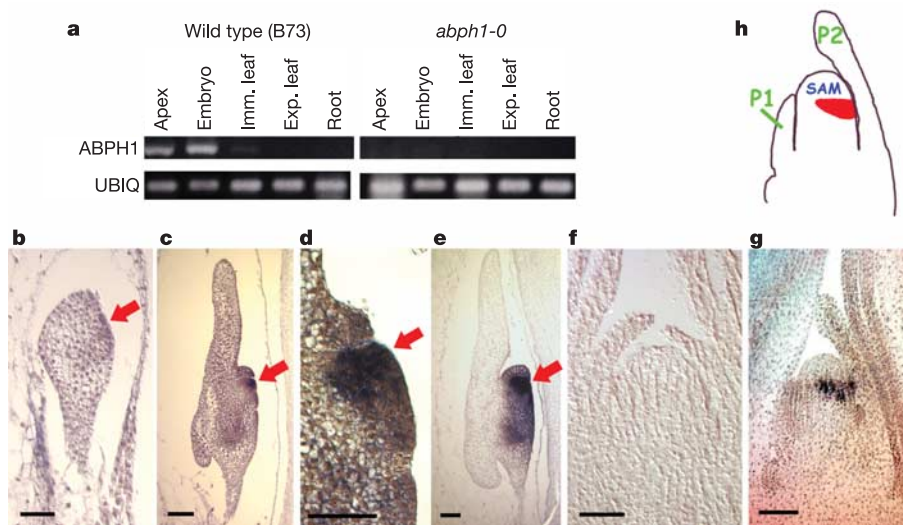


Figure 2 *ABPH1* gene expression. **a**, RT–PCR analysis shows strongest expression of *ABPH1* in shoot apex or embryo, very weak expression in immature (Imm.) leaf, and no expression in expanded (Exp.) leaf or in root. No expression was detected in *abph1-0* mutants. Ubiquitin (UBIQ) control amplifications show equal loading. **b**, *In situ* hybridization shows weak *ABPH1* expression in a transition stage embryo at the site of SAM initiation. **c–e**, Expression is stronger in the coleoptilar stage SAM (**c**), magnified in

(**d**), and persists in the first leaf stage SAM (**e**). Expression of *ABPH1* in **b–e** is marked by red arrows. **f–h**. In 2-week-old seedlings, no expression was detected in a control *abph1* mutant apex (**f**), and expression overlaps with the incipient leaf primordium in a normal apex (**g**), diagrammed in **h**. P1 and P2 indicate plastochron 1, 2 leaf. Scale bars: 50 μ m (**b, d**); 100 μ m (**c, e–g**).

of the maize and *Arabidopsis* A class response regulator gene families, suggesting that expansion of this gene family may have occurred after the monocotyledon–eudicotyledon divergence¹⁶.

We examined expression of *ABPH1* by RT–PCR. A transcript was present in shoot apex and embryo extracts, was very weak in immature leaf, and not detected in expanded leaf or root. No transcript was detected from the *abph1-0* allele (Fig. 2a). The development of *abph1* mutants first deviates from normal by initiation of a larger SAM in the embryo⁵. Using *in situ* hybridization, we first detected weak *ABPH1* expression at the site where the SAM forms on the flank of the transition stage embryo (Fig. 2b). By the coleoptilar stage (Fig. 2c, d), expression was stronger and again concentrated in the SAM region, and it persisted in the SAM at the first leaf stage (Fig. 2e). As expected, an *abph1* mutant apex had no hybridization signal (Fig. 2f). Later in development, in 2-week-old seedlings that had initiated around 10–15 leaves, *ABPH1* expression was restricted to a small domain of the SAM at the site of the incipient leaf primordium (Fig. 2f, g; see also Supplementary Fig. 2A). However, it did not label the entire incipient leaf primordium, which is marked by KN1 downregulation that fully encircles the SAM¹⁸. An *Arabidopsis* A class response regulator gene is also expressed in the SAM region and its expression correlates with SAM initiation in culture^{19,20}, suggesting that these genes may have similar functions. However, mutations in some of the *Arabidopsis* A class response regulators have relatively subtle effects on morphology, suggesting distinct functions or a higher degree of redundancy in *Arabidopsis* compared to maize¹³. *ABPH1* was also expressed in the female ear inflorescence¹⁶ (Supplementary Fig. 2B), although *abph1* mutants have no obvious phenotype in ears (not shown).

We next asked whether *ABPH1* was induced by cytokinin, as

suggested by homology^{21–23}. Seedling shoots, excised at the root–shoot junction, were incubated in water or in cytokinin for 30–120 min, then dissected into to a ‘shoot’ fraction (apical and axillary shoot meristems, stem and 4–5 young leaf primordia) or a ‘leaf’ fraction (expanding and mature leaves). *ABPH1* expression was induced in the shoot fraction within 30 min of cytokinin treatment, but not in the leaf fraction (Fig. 3a). *ABPH1* expression level was on average threefold and eightfold greater after 30 and 120 min cytokinin treatment, respectively, compared with water controls (mean of eight samples each, using from 10^{-4} M to 10^{-6} M kinetin or benzyladenine).

To investigate whether the localized pattern of *ABPH1* expression in the seedling SAM responded to cytokinin, we performed *in situ* hybridization on cytokinin-treated shoots. After 30 min cytokinin treatment, the *ABPH1* expression domain expanded to form a stripe across the entire SAM (Fig. 3b, compare to Fig. 2g). Longer cytokinin treatment led to additional induction of expression in young leaf primordia and vascular strands within the stem (Fig. 3c), although *ABPH1* expression was not induced in the central zone at the apex of the SAM. Although we do not fully understand the significance of these patterns, they suggest that endogenous *ABPH1* expression in the SAM is regulated either directly or indirectly by cytokinins.

The A class response regulators are thought to act as negative feedback regulators of cytokinin signalling^{8,9}. Therefore *ABPH1* may function by negatively regulating cytokinin signalling in the SAM. To test this hypothesis, we investigated whether we could phenocopy *abph1* mutants by culturing wild-type embryos in the presence of cytokinin. The youngest embryos that we could isolate were at the first leaf stage, and these were cultured in the presence or absence of cytokinin for 5 days, then fixed and cleared and the shoot meristems were measured. As expected, embryos grown on cytokinin showed reduced root growth (Fig. 3d, e)⁶. They also developed significantly larger shoot meristems, on average one-third wider than normal (Fig. 3f, g). The average SAM width for embryos grown without cytokinins was $100.3 \pm 5.9 \mu\text{m}$, and after growth on 10^{-6} or 10^{-7} M kinetin it was $134.7 \pm 12.6 \mu\text{m}$ ($n = 15$, difference significant at 0.1% level). Interestingly, this increase in SAM size is about the same as in *abph1* mutants⁵. We did not observe changes in phyllotaxy in cultured embryos, which may be because the youngest embryos that we could culture had already initiated the first leaf; thus their phyllotactic pattern was already established.

Our results indicate that cytokinin response regulator signalling controls the establishment of phyllotactic pattern in maize. These findings are consistent with studies linking SAM size with cytokinin concentration²⁴, and confirm a link between cytokinins, meristem size and phyllotaxy that was suggested by studies of *altered meristem programming1* mutants^{25,26}. As the defect in SAM size was phenocopied by culturing embryos on exogenous cytokinin, the simplest model for *ABPH1* function is that it acts in the embryo to restrict cytokinin-induced SAM proliferation. *ABPH1* therefore seems to regulate phyllotaxy by controlling the available signalling space in the meristem, consistent with repressor field theories of phyllotaxy or with biophysical models (reviewed in refs 4, 27–29). *ABPH1* may act upstream of the recently described auxin transport mechanism for determination of leaf position⁴, by controlling the dimensions of the morphogenetic domain in which the spatial coordinates of auxin transport are established. Later in development, localized *ABPH1* expression appears to overlap with that of the putative auxin efflux carrier PIN-FORMED1 at the incipient leaf initiation site⁴, suggesting that cytokinin signalling may also have a more specific role in determination of leaf position. This implies an interaction between cytokinin and auxin signalling, as already described for shoot regeneration⁶. Although future studies are required to test this hypothesis, the discovery of a role for cytokinin response regulation in phyllotaxy brings us a step closer to understanding the origin of these fascinating patterns. □

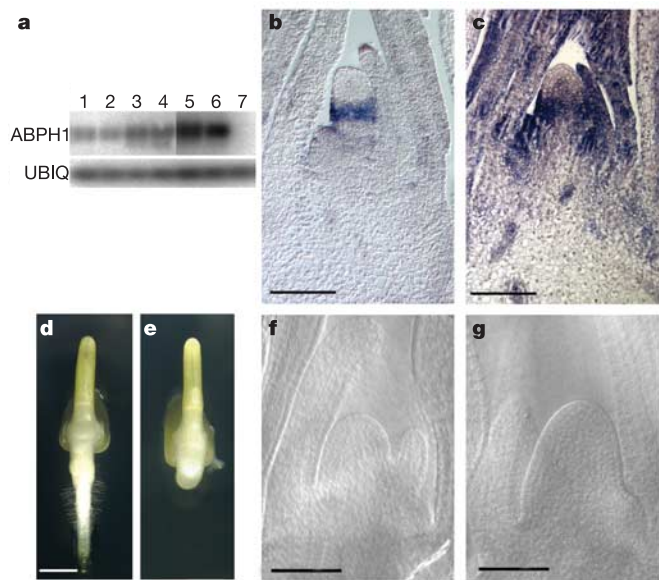


Figure 3 Effect of cytokinin on *ABPH1* expression and shoot meristem size. **a**, Southern blot of an RT–PCR gel of *ABPH1* expression in shoot apex at time zero or 120 min water control (lanes 1, 2), after 30 min treatment with 10^{-4} M or 10^{-5} M kinetin (lanes 3, 4), or after 120 min kinetin treatment (lanes 5, 6). *ABPH1* was not induced in leaf tissues even after 120 min cytokinin treatment (lane 7). The *UBIQUITIN* (*UBI*) control is also shown. **b, c**, In shoots treated with cytokinin, the expression domain of *ABPH1* mRNA expands after 30 min (**b**) or 120 min (**c**) treatment. Cytokinin also affects SAM development. **d, e**, Embryos excised at 16 days after pollination were cultured for 6 days in the absence (**d**) or presence (**e**) of 10^{-6} M kinetin. **f, g**, Clearing of the shoot apex revealed that the shoot meristem was bigger in cytokinin-treated embryos (**g**) compared with untreated embryos (**f**). Scale bar: 5 mm (**d, e**); 100 μm (**f, g**); 200 μm (**b, c**).

Methods

Plant materials

Plants were grown in the greenhouse or in the field under standard conditions. We used the B73 wild-type line. For cytokinin induction experiments, 2-week-old seedlings were excised at the shoot–root junction, and stood in water supplemented with different concentrations of cytokinin. After treatments, the seedlings were dissected into a shoot fraction (apical and axillary shoot meristems, stem and 4–5 young leaf primordia) or a leaf fraction (expanding and mature leaves). For embryo culture experiments, pollinated ears at either 13 days after pollination (first leaf stage) or 16 days after pollination (second to third leaf stage) were sterilized in 30% commercial bleach solution for 20 min then rinsed five times with sterile water. The embryos were dissected out and cultured on maize embryo culture medium (1 × Murashige and Skoog salts, 1 × Gamborg's vitamins, 2% sucrose, 0.7% agar, pH 5.7) in some cases with the addition of cytokinin (kinetin, 10^{-5} , 10^{-6} or 10^{-7} M). The embryos were cultured at 28 °C in the dark for 5–7 days then fixed and cleared and the shoot meristems were measured as described⁵.

Molecular biology

Standard protocols were used for maize DNA isolation and Southern blotting³⁰. For RT–PCR analysis, poly(A⁺) RNA was isolated using an Oligotex messenger RNA mini kit (Qiagen) according to the manufacturer's protocol. The primers ZmRR3f (GATGGCGAGCGCAAGTGT) and ZmRR3r (AATGCCGCTGCTACAGCTACCA) were used to amplify *ABPH1* transcripts using the one step RT–PCR kit (Qiagen). The control primers Ubi 5' (TAAGCTGCCGATGTGCTGCTGCTCG) and Ubi 3' (CTGAAAGACAGAACATAATGAGCACAG) were used to amplify control ubiquitin transcripts. PCR conditions were 94 °C, 15 s, 62 °C, 15 s, 72 °C, 45 s, with a 1 °C reduction in annealing temperature per cycle, touching down at 56 °C annealing, followed by 15 additional cycles. The PCR cycle number was limited to ensure semi-quantitative amplification, and no PCR product was visible on the ethidium-bromide-stained agarose gels. The gels were Southern blotted and probed with an *ABPH1* or *UBIQUITIN* probe. *ABPH1* transcript levels were estimated using a phosphorimager (Fuji) and were normalized against the respective ubiquitin values. *In situ* hybridizations were performed as described⁵.

Received 25 April; accepted 22 June 2004; doi:10.1038/nature02778.

- Jean, R. *Phyllotaxis: A Systemic Study in Plant Morphogenesis* (Cambridge Univ. Press, Cambridge, UK, 1994).
- Kuhlemeier, C. & Reinhardt, D. Auxin and phyllotaxis. *Trends Plant Sci.* **6**, 87–189 (2001).
- Klar, A. Fibonacci's flowers. *Nature* **417**, 595 (2002).
- Reinhardt, D. *et al.* Regulation of phyllotaxis by polar auxin transport. *Nature* **426**, 255–260 (2003).
- Jackson, D. & Hake, S. Control of phyllotaxy in maize by the *ABPHYL1* gene. *Development* **126**, 315–323 (1999).
- Skoog, F. & Miller, C. O. Chemical regulation of growth and organ formation in plant tissues cultured *in vitro*. *Symp. Soc. Exp. Biol.* **11**, 118–131 (1957).
- Mok, D. W. & Mok, M. C. Cytokinin metabolism and action. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **52**, 89–118 (2001).
- Sheen, J. Phosphorelay and transcription control in cytokinin signal transduction. *Science* **296**, 1650–1652 (2002).
- Hutchison, C. E. & Kieber, J. J. Cytokinin signaling in *Arabidopsis*. *Plant Cell* **14** (suppl.), S47–S59 (2002).
- Lohrmann, J. *et al.* The response regulator ARR2: a pollen-specific transcription factor involved in the expression of nuclear genes for components of mitochondrial complex I in *Arabidopsis*. *Mol. Genet. Genom.* **265**, 2–13 (2001).
- Hwang, I. & Sheen, J. Two-component circuitry in *Arabidopsis* cytokinin signal transduction. *Nature* **413**, 383–389 (2001).
- Sakai, H. *et al.* ARR1, a transcription factor for genes immediately responsive to cytokinins. *Science* **294**, 1519–1521 (2001).
- To, J. P. C. *et al.* Type-A *Arabidopsis* response regulators are partially redundant negative regulators of cytokinin signaling. *Plant Cell* **16**, 658–671 (2004).
- Greyson, R. I., Walden, D. B., Hume, J. A. & Erickson, R. O. The ABPHYL syndrome in *Zea mays*. II. Patterns of leaf initiation and the shape of the shoot meristem. *Can. J. Bot.* **56**, 1545–1550 (1978).
- Robertson, D., Stinard, P. & Maguire, M. Genetic evidence of Mutator-induced deletions in the short arm of chromosome 9 of maize. II. wd deletions. *Genetics* **136**, 1143–1149 (1994).
- Asakura, Y. *et al.* Molecular characterization of His-Asp phosphorelay signaling factors in maize leaves: Implications of the signal divergence by cytokinin-inducible response regulators in the cytosol and the nuclei. *Plant Mol. Biol.* **52**, 331–341 (2003).
- Falke, J., Bass, R., Butler, S., Chervinot, S. & Danielson, M. The two component signaling pathway of bacterial chemotaxis. *Annu. Rev. Cell Dev. Biol.* **13**, 437–512 (1997).
- Jackson, D., Veit, B. & Hake, S. Expression of maize *KNOTTED1* related homeobox genes in the shoot apical meristem predicts patterns of morphogenesis in the vegetative shoot. *Development* **120**, 405–413 (1994).
- D'Agostino, I. B., Deruere, J. & Kieber, J. J. Characterization of the response of the *Arabidopsis* response regulator gene family to cytokinin. *Plant Phys.* **124**, 1706–1717 (2000).
- Che, P., Gingerich, D. J., Lall, S. & Howell, S. H. Global and hormone-induced gene expression changes during shoot development in *Arabidopsis*. *Plant Cell* **14**, 2771–2785 (2002).
- Brandstatter, I. & Kieber, J. J. Two genes with similarity to bacterial response regulators are rapidly and specifically induced by cytokinin in *Arabidopsis*. *Plant Cell* **10**, 1009–1019 (1998).
- Sakakibara, H. *et al.* A response-regulator homologue possibly involved in nitrogen signal transduction mediated by cytokinin in maize. *Plant J.* **14**, 337–344 (1998).
- Taniguchi, M. *et al.* Expression of *Arabidopsis* response regulator homologs is induced by cytokinins and nitrate. *FEBS Lett.* **429**, 259–262 (1998).

- Werner, T., Motyka, V., Strnad, M. & Schmülling, T. Regulation of plant growth by cytokinin. *Proc. Natl Acad. Sci. USA* **98**, 10487–10492 (2001).
- Chaudhury, A. M., Letham, S., Craig, S. & Dennis, E. S. Amp1—a mutant with high cytokinin levels and altered embryonic pattern, faster vegetative growth, constitutive photomorphogenesis and precocious flowering. *Plant J.* **4**, 907–916 (1993).
- Helliwell, C. A. *et al.* The *Arabidopsis* AMP1 gene encodes a putative glutamate carboxypeptidase. *Plant Cell* **13**, 2115–2125 (2001).
- Schwabe, W. W. in *Positional Controls In Plant Development* (eds Barlow, P. W. & Carr, D. J.) (Cambridge Univ. Press, 1984).
- Callos, J. D. & Medford, J. I. Organ positions and pattern formation in the shoot apex. *Plant J.* **6**, 1–7 (1994).
- Green, P. B. Connecting gene and hormone action to form, pattern and organogenesis; biophysical transductions. *J. Exp. Bot.* **45**, 1775–1788 (1994).
- Taguchi-Shiobara, F., Yuan, Z., Hake, S. & Jackson, D. The *FASCIATED EAR2* gene encodes a leucine-rich repeat receptor-like protein that regulates shoot meristem proliferation in maize. *Genes Dev.* **15**, 2755–2766 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank V. Chandler for *Spm* transposon lines, and members of the Jackson laboratory, C. Kidner, E. Vollbrecht and P. Sherwood, for comments on the manuscript. We also thank Z. Yuan and M. Krishnaswami for assistance with genetic screens, DNA isolations and Southern blotting, and T. Mulligan for help with plant propagation. Funding from the National Science Foundation (Plant and Animal Developmental Mechanisms) is also acknowledged.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.J. (jackson@csh.ledu).

Suppression of anoikis and induction of metastasis by the neurotrophic receptor TrkB

Sirith Douma¹, Theo van Laar¹, John Zevenhoven¹, Ralph Meuwissen¹, Evert van Garderen² & Daniel S. Peepers¹

¹Division of Molecular Genetics, and ²Department of Experimental Animal Pathology, The Netherlands Cancer Institute, Plesmanlaan 121, Amsterdam 1066 CX, The Netherlands

Metastasis is a major factor in the malignancy of cancers, and is often responsible for the failure of cancer treatment. Anoikis (apoptosis resulting from loss of cell–matrix interactions) has been suggested to act as a physiological barrier to metastasis; resistance to anoikis may allow survival of cancer cells during systemic circulation, thereby facilitating secondary tumour formation in distant organs^{1–3}. In an attempt to identify metastasis-associated oncogenes, we designed an unbiased, genome-wide functional screen solely on the basis of anoikis suppression. Here, we report the identification of TrkB, a neurotrophic tyrosine kinase receptor^{4,5}, as a potent and specific suppressor of caspase-associated anoikis of non-malignant epithelial cells. By activating the phosphatidylinositol-3-OH kinase/protein kinase B pathway, TrkB induced the formation of large cellular aggregates that survive and proliferate in suspension. In mice, these cells formed rapidly growing tumours that infiltrated lymphatics and blood vessels to colonize distant organs. Consistent with the ability of TrkB to suppress anoikis, metastases—whether small vessel infiltrates or large tumour nodules—contained very few apoptotic cells. These observations demonstrate the potent oncogenic effects of TrkB and uncover a specific pro-survival function that may contribute to its metastatic capacity, providing a possible explanation for the aggressive nature of human tumours that overexpress TrkB.

To perform a functional screen for suppressors of anoikis, we selected rat intestinal epithelial (RIE) cells owing to their non-malignancy and high sensitivity to anoikis (refs 6, 7; see below). Anoikis can be induced *in vitro* by transferring epithelial cells from standard, adhesive cell culture dishes (with a hydrophilic surface that supports cell attachment and spreading) to ultra-low cluster (ULC) plates with a covalently bound hydrogel layer that effectively inhibits cellular attachment. Shifting RIE cells from adhesive to ULC dishes resulted in massive cell death (Fig. 1a, left), which was critical in view of the desired minimal number of surviving 'background' cells that would represent false positives in the screen. In agreement with previous results⁷, oncogenic Ras^{V12} protein effectively suppressed anoikis (Fig. 1a, right).

Upon retroviral delivery of a high-complexity pEYK retroviral complementary DNA library⁸, cells were brought into suspension and screened for surviving clones (Fig. 1b). Relative to control green fluorescent protein (GFP)-infected cells that died, a proportion of library-infected pools showed surviving clones (see Methods), including 'ME-3'; this clone effectively suppressed anoikis and formed large, spheroid aggregates in suspension (Fig. 1c). Isolation and sequencing of its proviral insert revealed a full-length, wild-type cDNA encoding TrkB, a neurotrophic tyrosine kinase receptor that is required for the ontogeny, function and survival of the mammalian nervous system^{4,5} and other cell types^{9–12}. Interestingly, neurite development and outgrowth may represent processes that particu-

larly rely on adhesion-independent survival¹³.

To characterize the biological effects of TrkB, we generated, by retroviral delivery, pools of epithelial cells stably expressing TrkB, its primary ligand (brain-derived neurotrophic factor (BDNF))^{14,15}, or both (Supplementary Fig. S1a). On adhesive plates, TrkB, but not BDNF, disrupted epithelial cell organization, which was aggravated by ligand co-expression, resulting in complete loss of cell-cell contact and, upon reaching confluency, in continued proliferation as spheroids in suspension (Fig. 1d, Supplementary Fig. S1b, c).

To confirm that TrkB suppresses anoikis, we followed the fate of TrkB-expressing epithelial cells in suspension. In ULC dishes, both empty-vector and BDNF-expressing cells underwent cell death, starting at day two and leaving no surviving cells after two weeks (Fig. 1e). By contrast, TrkB effectively suppressed cell death and allowed cells to proliferate as large spheroid aggregates in suspension, explaining the identification of TrkB in the screen. Co-expression of TrkB with BDNF led to a further increase in cell number (Supplementary Fig. S1d). Similar results were obtained for murine and human epithelial cells (Supplementary Fig. S1e). The observation that the entire polyclonal TrkB-expressing cell population was resistant to anoikis argued against clonal events being due to genetic mutations. This was further supported by a soft-agar assay, which reinforced our notion that TrkB expression bypasses the need for anchorage, with virtually every TrkB-expressing cell producing a colony (Fig. 1f).

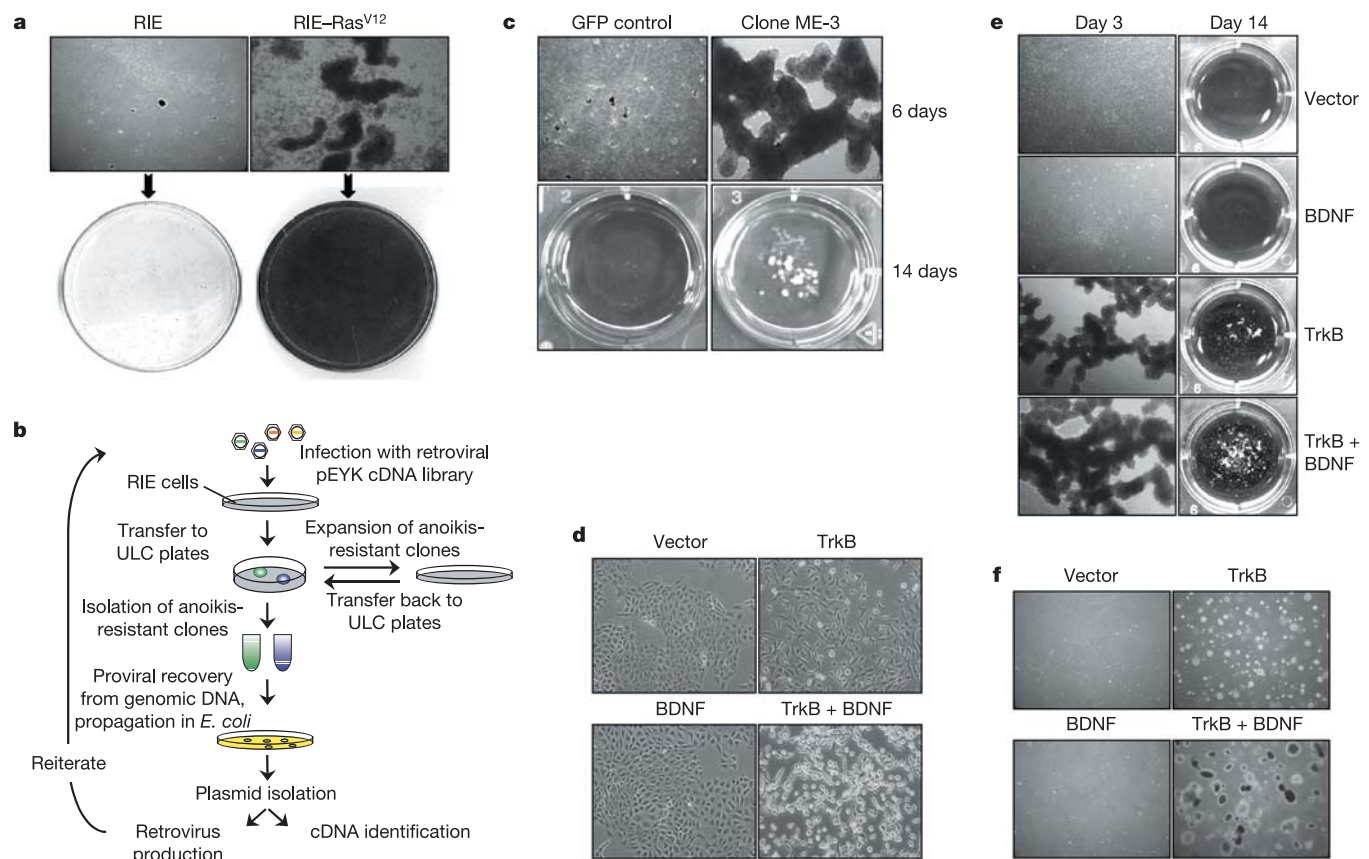


Figure 1 Functional cloning and validation of TrkB as a suppressor of anoikis. **a**, Parental or Ras^{V12}-expressing RIE cells seeded into ULC tissue culture plates and photographed at $\times 50$ magnification after 3 days (top). The bottom panel shows subsequent transfer of the cells to adhesive tissue culture plates, followed by fixation and crystal violet staining 2 days later. **b**, Outline of the screen. See Methods for details. **c**, Suppression of anoikis by clone ME-3, relative to a GFP-control, photographed at $\times 50$ (top) and $\times 1$ (bottom)

magnification and at indicated time points. **d**, RIE cells expressing cDNAs as indicated, seeded on adhesive plates and photographed at $\times 200$ magnification. **e**, RIE cells expressing cDNAs as indicated, seeded into ULC plates and photographed in time at $\times 50$ (day 3) and $\times 1$ (day 14) magnification. **f**, Single cell suspensions of RIE cells expressing cDNAs as indicated, seeded into soft-agar and photographed at $\times 50$ magnification 1 week later.

As anoikis is defined as apoptosis caused by lack of adhesion², we examined whether it was specifically apoptosis that was prevented by TrkB. First, staining with an apoptosis-dependent fluorescent dye (3,3'-dihexyloxa-carbocyanine iodide, DiOC₆(3)) showed that, in agreement with previous results^{6,7}, RIE cells underwent apoptosis upon attachment withdrawal (Fig. 2a, left). By contrast, apoptosis was completely blocked by TrkB (Fig. 2a, right). Second, we determined by immunoblotting the protein levels of cleaved caspase-3, a primary executioner of apoptosis. Whereas no activated caspase-3 was detected in any of the cell lines when grown on adhesive dishes, it was produced by control and BDNF-expressing cells when they were brought into suspension (Fig. 2b). By contrast, cleaved caspase-3 production was completely suppressed by TrkB, regardless of the presence of TrkB's ligand. Together, these results indicate that it is apoptosis, resulting from lack of adhesion, that is effectively suppressed by TrkB.

Integrins are principal mediators of adhesion between cells and extracellular matrix proteins, including fibronectin, and they transduce signals critically required for cell survival^{16,17}. Therefore, we examined whether TrkB-mediated resistance to anoikis depends on fibronectin, or on any other potential survival factor(s) present in cell culture serum. When TrkB-expressing cells were deprived of serum, they were partially protected from apoptosis, relative to control cells (Fig. 2c). However, BDNF-stimulated TrkB rendered cells resistant to anoikis, despite the complete absence of serum. This effect was not due to a general pro-survival function of TrkB, as TrkB, in contrast to the anti-apoptotic Bcl-2 protein, completely failed to protect adherent c-Myc-expressing fibroblasts from apoptosis resulting from serum withdrawal (Fig. 2d, e). Therefore, in untransformed epithelial cells, the BDNF-TrkB axis generates a signal that is sufficiently potent to allow, in the absence of any exogenous

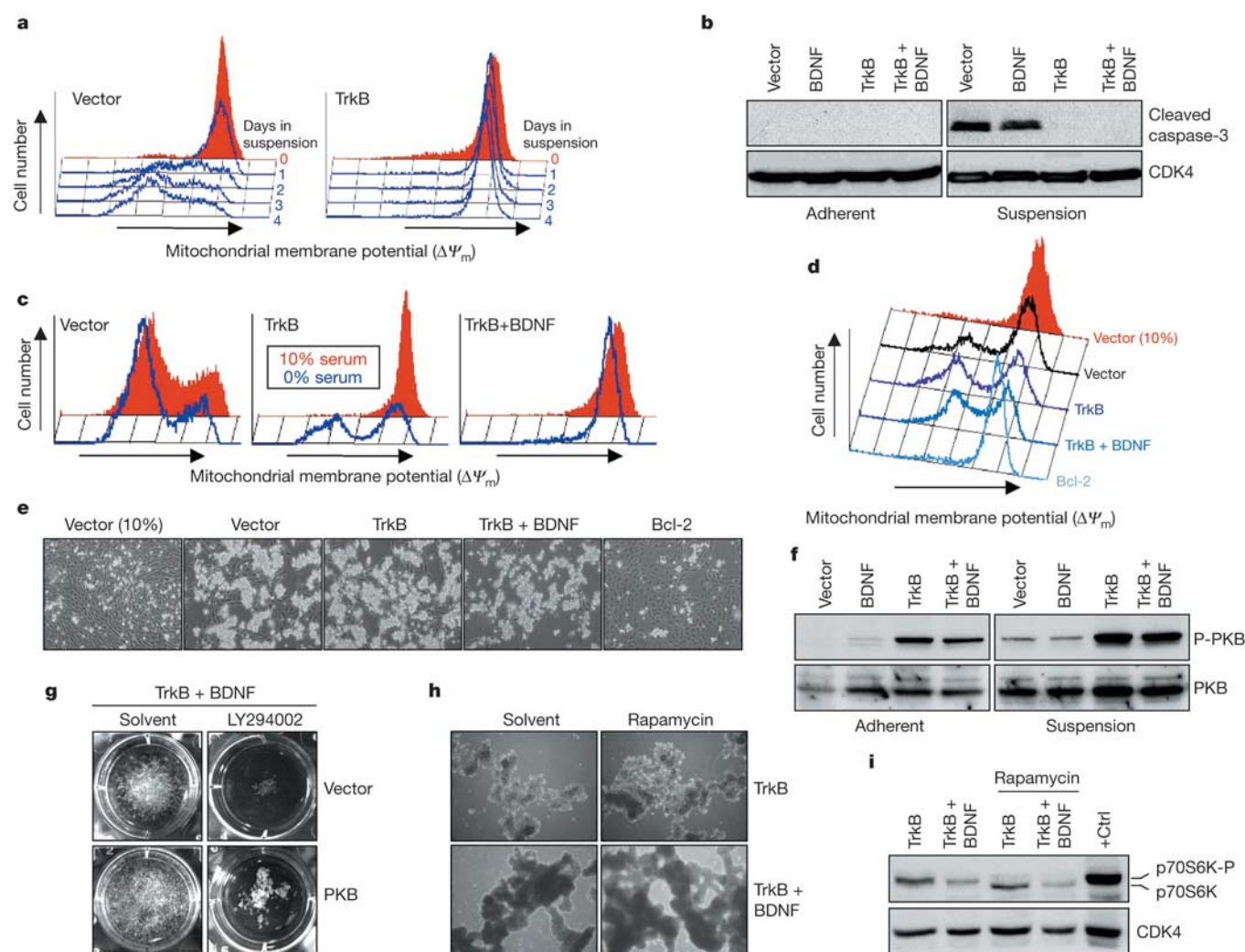


Figure 2 Specific suppression of anoikis-associated apoptosis by TrkB. **a**, Control or TrkB-expressing RIE cells seeded into ULC plates and analysed for apoptosis over time by DiOC₆(3) staining for FACS analysis³⁰. Right- and left-hand peaks reflect living and apoptotic cells respectively. **b**, Immunoblotting for cleaved caspase-3 of lysates from RIE cells expressing indicated cDNAs, on adhesive (left) or ULC (right) plates, for 2 days. CDK4 serves as loading control. **c**, RIE cells expressing indicated cDNAs, seeded either in the presence (red) or absence (blue) of serum into ULC plates and analysed for apoptosis 3 days later, as in **a**. Results with 0.1% serum were similar to those with 0% (not shown). **d**, **e**, Rat1 fibroblasts co-expressing c-Myc and indicated cDNAs, on adhesive dishes, in the absence of serum for 3 days and analysed by FACS (**d**) and photographed at $\times 200$

magnification (**e**). Empty vector-transduced cells in 10% serum (Vector 10%) serve as control. **f**, Immunoblotting for serine 473-phosphorylated PKB (P-PKB) and total PKB of lysates from RIE cells expressing indicated cDNAs, on adhesive (left) or ULC plates (right) for 2 days. **g**, TrkB/BDNF co-expressing cells, either in the presence or absence of activated PKB, treated after spheroid formation with LY294002 (20 μ M) or solvent and photographed 12 days later at $\times 1$ magnification. **h**, **i**, TrkB/BDNF-expressing cells treated after spheroid formation with rapamycin (20 nM), photographed at $\times 50$ magnification 2 days later (**h**) and subsequently used for immunoblotting for p70S6K and CDK4 as loading control (**i**). Similar results were obtained for 1-week treatment (not shown).

factors, survival in the face of a strong and specific pro-apoptotic anoikis signal.

To investigate whether TrkB directly activates survival signalling, as well as to identify its downstream targets, we first examined the potential involvement of protein kinase B (PKB; also known as AKT), which by itself is sufficient to suppress anoikis^{18,19}. Under adherent conditions, TrkB induced production of activated, phosphorylated PKB (Fig. 2f, left). As TrkB/BDNF-expressing cells on adhesive dishes show diminished cell–cell interactions (Fig. 1d), this result demonstrates that, rather than through promoting cell–cell adhesion, TrkB directly activates the PKB survival pathway. This result, however, does not rule out the possibility that aggregate formation may contribute to survival signalling in suspension. Importantly, PKB activation was maintained in cells challenged with anoikis (Fig. 2f, right), raising the possibility that the PKB pathway mediates anoikis suppression by TrkB. Indeed, blocking PKB-dependent signalling by pharmacological inhibition of its major upstream regulator, phosphatidylinositol-3-OH kinase (PI(3)K), with LY294002 significantly interfered with TrkB-mediated survival (Fig. 2g; Supplementary Fig. S2a, b). In turn, cell death was prevented by activated PKB, excluding aspecific cytotoxic effects of LY294002, and further supporting a central role for PKB in anoikis suppression by TrkB. In line with previous data on Madine–Darby canine kidney (MDCK) cells¹⁸, PI(3)K was also sufficient to suppress anoikis, albeit to a lesser degree than TrkB (Supplementary Fig. S2c). One of the critical downstream targets of the PI(3)K/PKB pathway is p70S6K, a kinase implicated both in anoikis and lymphomagenesis^{18,20}. However, although rapamycin (a specific inhibitor of mTOR, which is a regulator of p70S6K) blocked p70S6K phosphorylation, it completely failed to affect the survival of TrkB-expressing cells in suspension (Fig. 2h, i). Similarly, dominant interference with signalling by Rac proteins (another target of PKB), or pharmacological inhibition of MEK did not prevent suppression of anoikis by TrkB (Supplementary Fig. S2d, e). We conclude that TrkB activates PI(3)K/PKB signalling, which contributes to anoikis resistance.

The main objective of this study was to assess the feasibility of using anoikis suppression as a basis to identify metastasis-

inducing (onco)genes. When non-malignant cells expressing TrkB, either alone or in conjunction with its ligand, were introduced into nude mice, they rapidly formed tumours with high efficiencies, whereas vector and BDNF-expressing cells failed to do so (Supplementary Table S1). Therefore, expression of TrkB is sufficient to convert non-malignant epithelial cells into highly tumorigenic cells.

To address whether TrkB endows epithelial cells with metastatic potential as well, we first re-engineered our cell panel to express luciferase, which allows non-invasive *in vivo* imaging in live mice²¹. Each polyclonal cell line was administered intravenously to nude mice, and outgrowth and the location of tumours was monitored. Luciferase imaging showed that, although at 90 min after injection control cells were apparently viable and had accumulated in the lungs to at least the same extent as TrkB/BDNF co-expressing cells, no luciferase signal was detectable at 6 days (Fig. 3a). Indeed, control and BDNF-expressing cells failed to form tumours within 100 days after inoculation (Fig. 3b), consistent with the hypothesis that non-malignant epithelial cells fail to survive for a prolonged period in the absence of appropriate attachment². By contrast, TrkB-expressing epithelial cells colonized the lungs and heart, where they formed rapidly growing tumours (Fig. 3c). This phenotype was even more pronounced for TrkB/BDNF-expressing cells, which caused metastases throughout the body (Fig. 3a, c; Supplementary Fig. S3). The cooperation between TrkB and BDNF in inducing metastasis therefore recapitulated their collaboration in suppressing anoikis under low-serum conditions (which may better approximate the situation *in vivo* than does 10% serum; compare Fig. 2c to Fig. 3c).

Microscopic pathological analysis showed the capacity of TrkB to induce solid, non-encapsulated tumours with undifferentiated properties (Fig. 4 and Supplementary Fig. S4). The tumours showed malignant cellular characteristics, including a high mitotic index, a high nucleus:cytoplasm ratio, and notable cellular and nuclear pleomorphism. TrkB-specific immunohistochemistry showed that these tumours were able to efficiently invade both blood and lymphatic vessels in various tissues, including liver, kidney, lung and heart (Fig. 4a–f, i, j and Supplementary Fig. S4). This infiltrative character of TrkB-driven tumours was also markedly illustrated by

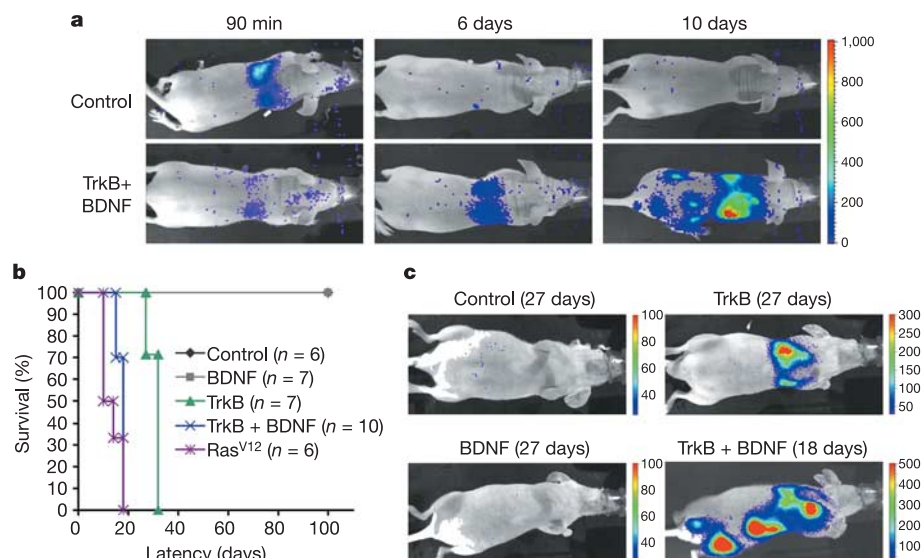


Figure 3 Tumorigenic potential of TrkB. **a–c**, Mice injected intravenously with RIE cells co-expressing a luciferase gene and indicated cDNAs, and subjected to *in vivo* imaging at indicated time points, with relative light units per pixel indicated in colour (**a**, **c**), or

represented in a Kaplan–Meier survival plot (**b**). Note the relatively short latency of TrkB/BDNF-induced tumours and different imaging sensitivities used (**c**).

their capacity to destruct and invade bones (Supplementary Fig. S4g–i). Furthermore, these tumours, when present as large tumour masses, were viable and proliferating, as judged by their very low apoptotic and high proliferative indices respectively (Fig. 4g, h, k). Importantly, this viable and non-apoptotic character was retained when these tumours extended into vessels (Fig. 4i, j), conceivably contributing to their metastatic potential and consistent with our finding *in vitro* that TrkB suppresses anoikis. The tumours were so invasive, destructive and fast-growing that only 18 days after administration, all mice had been euthanized (Fig. 3b, Supplementary Fig. S3).

As the intravenous metastasis system precludes an assessment of the early stages of metastasis (tissue invasion and endothelial or lymphatic intravasation by disseminated tumour cells), we also introduced TrkB/BDNF co-expressing cells subcutaneously. To increase the time window in which metastasis could occur, tumours that arose within a week were surgically removed, and upon relapse mice were killed. In agreement with, and extending the results described above, pathological examination showed that the subcutaneous tumour cells had efficiently infiltrated muscles, lymphatic vessels and regional lymph nodes, eventually causing several

metastases at the pleural side of the lungs (Fig. 4l–p). Together, these results demonstrate that TrkB acts as a potent oncogene that converts non-malignant epithelial cells into metastasizing tumour cells.

TrkB and BDNF are frequently overexpressed in human cancer, including pancreatic and prostate carcinomas, Wilms' tumour and neuroblastomas, particularly those with aggressive behaviour and poor prognosis^{22–25}. Moreover, recent sequence analysis of the tyrosine kinase in colorectal cancers has shown mutations within the kinase domain of TrkB²⁶. Although these observations have associated TrkB with cancer, it has been unclear whether deregulated TrkB can act oncogenically. We find that TrkB generates a specific and potent pro-survival signal that renders epithelial cells resistant to anoikis. This is accompanied by the acquirement of potent tumorigenic, invasive and metastatic capacities, arguing that TrkB may directly contribute to human malignancies. Although our data are consistent with the hypothesis that suppression of anoikis contributes to the acquisition of a metastatic phenotype, other properties of TrkB, including induction of invasiveness, may contribute to its aggressive tumorigenic effects.

The propensity to metastasize may be determined by the gene

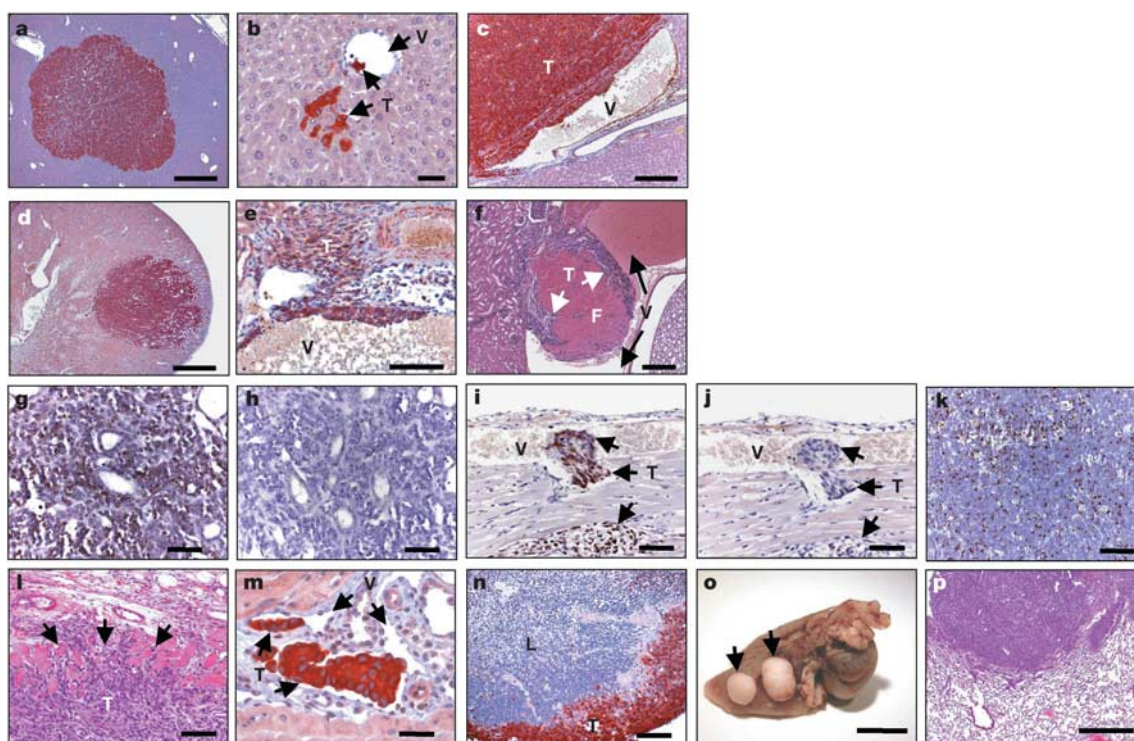


Figure 4 Metastatic potential of TrkB. **a–j**, Pathological characteristics of tumour processes after tail vein injection into nude mice. **a**, Large tumour nodule within liver. **b**, Clusters of tumour cells located in liver sinusoids and a tumour cell located in vena hepatica branch. **c**, Tumour nodule within liver, invading a vena portae branch. **d**, Large tumour within kidney. **e**, Detail of a renal tumour growing within lumen of large vein. **f**, Another plane of the section shown in **e**, demonstrating renal tumour extending into vein, activating the blood coagulation system and causing fibrin deposition (F). **g**, Detail of tumour in **d**, showing tumour cells located between renal tubules, and high proliferative index, as evident from nuclear MCM7 immunoreactivity. **h**, Representative photomicrograph of section in **g**, showing lack of apoptosis, as evident from absence of cleaved caspase-3 immunoreactivity. **i**, Tumour nodule (with edge of large tumour mass at bottom side of microphotograph), spreading into branch of coronary vein of the myocardium and showing high proliferative index, as evident from MCM7 immunopositivity. **j**, Representative photomicrograph of adjacent section of **i**, showing lack of apoptosis of tumour cells infiltrating coronary vein, as evidenced by a lack of

expression of cleaved caspase-3. **k**, Splenic lymphoma from an unrelated mouse, showing numerous apoptotic cells, serving as positive control for cleaved caspase-3-specific immunohistochemistry. **l–p**, Pathological characteristics of tumour processes upon subcutaneous injection. Representative microphotographs showing the metastatic process in one mouse. **l**, Subcutaneously injected tumour cells infiltrating muscle tissue (arrows). **m**, Clusters of tumour cells infiltrating lymphatic vessels, at a distance from the subcutaneous tumour. **n**, Tumour infiltrating axillary lymph node (L) draining the tumour site. **o**, Large metastases at pleural side of the lungs (apical lobe of the right lung was removed to allow a better view of the tumour nodules). **p**, One of the metastases shown in **o**. Immunohistochemistry was performed for TrkB (**a–e**, **m**, **n**), MCM7 (**g**, **i**), or cleaved caspase-3 (**h**, **j**, **k**). **f**, **l**, **p**, Haematoxylin and eosin stainings. T and V mark tumours and vessels, respectively. Scale bars: 500 µm (**a**, **p**); 200 µm (**c**, **e**); 100 µm (**f**, **k**, **l**, **n**); 50 µm (**b**, **g–j**); 20 µm (**m**); 1 mm (**d**); 5 mm (**o**). Rows indicate four individual sets of photographs.

expression signature already present in the primary tumour²⁷. The results presented here raise the possibility that TrkB activation may represent such an early event during multistep tumorigenesis, endowing primary cells with a wide spectrum of capacities contributing to cancer. Mainly on the basis of the overexpression of TrkB in poor-prognosis human tumours, pan-Trk-inactivating drugs are already being developed²⁸. Our findings predict that drug-mediated inactivation of specifically TrkB may have a notable impact on the tumorigenic and metastatic capacities of human tumours with overexpressed or mutated TrkB. Finally, our results not only illustrate the feasibility of using ankois suppression as a functional property to identify metastasis-associated oncogenes, but also show its potential use in high-throughput screens for TrkB-inhibitory cancer therapeutics. □

Methods

Genome-wide retroviral cDNA screen

RIE cells were infected, in the presence of $5 \mu\text{g ml}^{-1}$ polybrene, with a high-complexity retroviral pEYK cDNA library⁸, prepared from a whole mouse embryo and propagated in Phoenix packaging cells as eight pools representing two size fractions (1–3 and >3 kilobases (kb)). Filtered (0.45 μm) viral supernatant pools representing the two size fractions were used to infect 0.75×10^6 cells each, and 2 days later, cells were collected by trypsinization and transferred to four ULC six-well cell culture dishes (Costar) in total. After 6 days, cells were transferred back to regular, adhesive tissue culture dishes, to allow ankois-resistant cells to expand. After 10 days, cells were re-challenged with an ankois insult, by transferring them into ULC plates. One week later, surviving colonies were isolated and transferred to 24-well adhesive plates, and allowed to expand. Then, genomic DNA was isolated and digested with either *NotI* or *AscI*, re-ligated and transformed into *Escherichia coli*. Plasmid DNA was isolated and transfected onto Phoenix cells. Virus supernatant was subsequently used to infect fresh RIE cells and a second-round screen was performed to validate the biological activity of the respective proviral cDNAs. Finally, proviral cDNA inserts were isolated from positive clones, shuttled and analysed by DNA sequencing. One of the ankois-resistant clones (ME-3) contained a wild-type, full-length murine cDNA encoding TrkB, which was sub-cloned into the pBabe-puro retroviral vector. Of the 24 wells screened, 12 contained ankois-resistant clones, sometimes several per well, giving 20 ankois-resistant clones in total. Polymerase chain reaction (PCR) analysis for retroviral TrkB inserts showed that, in addition to ME-3, TrkB was present in seven clones (derived from four different wells). The remaining 13 clones remain to be analysed further.

A murine full-length cDNA encoding BDNF was isolated from a mouse cDNA library (Clontech) by PCR (primers available upon request), and sub-cloned into the pBabe-hygro retroviral vector.

Cell culture and *in vitro* assays

RIE cells were maintained in high-glucose Dulbecco's modified Eagle's Medium (DMEM) (Life Technologies) supplemented with 9% fetal bovine serum (PAA Laboratories) and antibiotics. c-Myc-expressing Rat1 fibroblasts were provided by K. Berns. Phoenix packaging cells were used to generate ecotropic retroviruses, as described²⁹. In general, a single-infection round of 6 h was sufficient to infect at least 90% of the population. RIE cells were infected with pBabe retroviral vectors carrying selectable markers and cDNA inserts, and at 1 to 2 days post infection selected with puromycin ($1.5 \mu\text{g ml}^{-1}$) or hygromycin ($75 \mu\text{g ml}^{-1}$) for at least 5 days. After confirming that all mock-infected cells were dead, the cells were used in various assays.

Proliferation curves were performed with two independent cultures for each cell line, and data points were collected in triplicate, as described²⁹.

To determine the cloning efficiency of RIE cell lines in semi-solid medium, soft agar assays were performed, as described²⁹.

To measure apoptosis, cells were collected from either adhesive or suspension plates (both populations were subjected to trypsinization), treated with DiOC₆(3) (a fluorescing dye that accumulates in mitochondria as a function of membrane potential³⁰; Molecular Probes) in PBS for 15 min at 37 °C, and subsequently analysed by fluorescence-activated cell sorting (FACS).

Pharmacological inhibition was performed with LY294002 (20 μM), PD98059 (20 μM), K-252A (300 nM) or rapamycin (20 nM), all from Calbiochem. Retroviral vectors encoding myristylated PKB and activated p110-CAAX PI(3)K were gifts from F. Scheeren and J. Collard respectively.

Immunohistochemistry

Paraffin-embedded tissue was cut into 4- μm sections, deparaffinized and rehydrated. Upon antigen retrieval performed with citrate buffer (pH, 6) and incubation with primary and secondary antibodies, immunocomplexes were detected with either DAB/H₂O₂ or AEC (Dakocytomation).

Antibodies

For western blotting and immunohistochemistry, the following antibodies were used: MS-862 for MCM-7 (NeoMarkers); AC-74 for actin (Sigma); 9661 for cleaved caspase-3 and 9202 for p70S6K (Cell Signalling); SC-8316 for TrkB, SC-260 for CDK4, SC-7985 for Ser 473-phosphorylated PKB and SC-8312 for PKB (all Santa Cruz).

In vivo assays

To determine tumorigenicity and metastatic potential, the RIE cell panel (1×10^6 cells) was injected subcutaneously into athymic nude mice in each flank. Mice were inspected daily and were killed at the time tumours reached a diameter of 10 mm.

To measure metastatic potential and allow *in-vivo* bioluminescence luciferase imaging, mice were injected intravenously with the RIE cell panel (1×10^6 for Fig. 3b, c, and 3×10^6 for Fig. 3a; administration of 10^7 cells gave identical results; not shown), co-expressing a luciferase gene and cDNAs of interest. *In vivo* bioluminescence imaging of luciferase expression was performed as described²¹. Generally, for each cell line, 2–4 mice were imaged in time, with similar results.

Received 27 April; accepted 16 June 2004; doi:10.1038/nature02765.

1. Frisch, S. M. & Ruoslahti, E. Integrins and ankois. *Curr. Opin. Cell Biol.* **9**, 701–706 (1997).
2. Frisch, S. M. & Sreton, R. A. Ankois mechanisms. *Curr. Opin. Cell Biol.* **13**, 555–562 (2001).
3. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
4. Klein, R., Parada, L. F., Coulier, F. & Barbacid, M. trkB, a novel tyrosine protein kinase receptor expressed during mouse neural development. *EMBO J.* **8**, 3701–3709 (1989).
5. Huang, E. J. & Reichardt, L. F. Trk receptors: roles in neuronal signal transduction. *Annu. Rev. Biochem.* **72**, 609–642 (2003).
6. Shao, J., Sheng, H., DuBois, R. N. & Beauchamp, R. D. Oncogenic Ras-mediated cell growth arrest and apoptosis are associated with increased ubiquitin-dependent cyclin D1 degradation. *J. Biol. Chem.* **275**, 22916–22924 (2000).
7. McFall, A. *et al.* Oncogenic Ras blocks ankois by activation of a novel effector pathway independent of phosphatidylinositol 3-kinase. *Mol. Cell Biol.* **21**, 5488–5499 (2001).
8. Koh, E. Y., Chen, T. & Daley, G. Q. Novel retroviral vectors to facilitate expression screens in mammalian cells. *Nucleic Acids Res.* **30**, e142 (2002).
9. Glass, D. J. *et al.* TrkB mediates BDNF/NT-3-dependent survival and proliferation in fibroblasts lacking the low affinity NGF receptor. *Cell* **66**, 405–413 (1991).
10. Jaboin, J., Hong, A., Kim, C. J. & Thiele, C. J. Cisplatin-induced cytotoxicity is blocked by brain-derived neurotrophic factor activation of TrkB signal transduction path in neuroblastoma. *Cancer Lett.* **193**, 109–114 (2003).
11. Ho, R. *et al.* Resistance to chemotherapy mediated by TrkB in neuroblastomas. *Cancer Res.* **62**, 6462–6466 (2002).
12. Klein, R. *et al.* Targeted disruption of the trkB neurotrophin receptor gene results in nervous system lesions and neonatal death. *Cell* **75**, 113–122 (1993).
13. Banks, G. B. & Noakes, P. G. Elucidating the molecular mechanisms that underlie the target control of motoneuron death. *Int. J. Dev. Biol.* **46**, 551–558 (2002).
14. Soppet, D. *et al.* The neurotrophic factors brain-derived neurotrophic factor and neurotrophin-3 are ligands for the trkB tyrosine kinase receptor. *Cell* **65**, 895–903 (1991).
15. Squinto, S. P. *et al.* trkB encodes a functional receptor for brain-derived neurotrophic factor and neurotrophin-3 but not nerve growth factor. *Cell* **65**, 885–893 (1991).
16. Ruoslahti, E. & Vaheri, A. Cell-to-cell contact and extracellular matrix. *Curr. Opin. Cell Biol.* **9**, 605–607 (1997).
17. Reginato, M. J. *et al.* Integrins and EGFR coordinately regulate the pro-apoptotic protein Bim to prevent ankois. *Nature Cell Biol.* **5**, 733–740 (2003).
18. Khwaja, A., Rodriguez-Viciana, P., Wennstrom, S., Warne, P. H. & Downward, J. Matrix adhesion and Ras transformation both activate a phosphoinositide 3-OH kinase and protein kinase B/Akt cellular survival pathway. *EMBO J.* **16**, 2783–2793 (1997).
19. Schmidt, M., Hovellmann, S. & Beckers, T. L. A novel form of constitutively active farnesylated Akt1 prevents mammary epithelial cells from ankois and suppresses chemotherapy-induced apoptosis. *Br. J. Cancer* **87**, 924–932 (2002).
20. Wendel, H. G. *et al.* Survival signalling by Akt and eIF4E in oncogenesis and cancer therapy. *Nature* **428**, 332–337 (2004).
21. Vooijs, M., Jonkers, J., Lyons, S. & Berns, A. Noninvasive imaging of spontaneous retinoblastoma pathway-dependent tumours in mice. *Cancer Res.* **62**, 1862–1867 (2002).
22. Nakagawa, A., Azar, C. G., Scavarda, N. J. & Brodeur, G. M. Expression and function of TRK-B and BDNF in human neuroblastomas. *Mol. Cell Biol.* **14**, 759–767 (1994).
23. Aoyama, M. *et al.* Human neuroblastomas with unfavorable biologies express high levels of brain-derived neurotrophic factor mRNA and a variety of its variants. *Cancer Lett.* **164**, 51–60 (2001).
24. Eggert, A. *et al.* Expression of the neurotrophin receptor TrkB is associated with unfavorable outcome in Wilms' tumor. *J. Clin. Oncol.* **19**, 689–696 (2001).
25. Brodeur, G. M. Neuroblastoma: biological insights into a clinical enigma. *Nature Rev. Cancer* **3**, 203–216 (2003).
26. Bardelli, A. *et al.* Mutational analysis of the tyrosine kinase in colorectal cancers. *Science* **300**, 949 (2003).
27. Bernards, R. & Weinberg, R. A. A progression puzzle. *Nature* **418**, 823 (2002).
28. Ruggeri, B. A., Miknyoczki, S. J., Singh, J. & Hudkins, R. L. Role of neurotrophin-trk interactions in oncology: the anti-tumor efficacy of potent and selective trk tyrosine kinase inhibitors in pre-clinical tumor models. *Curr. Med. Chem.* **6**, 845–857 (1999).
29. Peeper, D. S. *et al.* A functional screen identifies hDRIL1 as an oncogene that rescues RAS-induced senescence. *Nature Cell Biol.* **4**, 148–153 (2002).
30. Zamzami, N. *et al.* Mitochondrial control of nuclear apoptosis. *J. Exp. Med.* **183**, 1533–1544 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. van der Valk and J.-Y. Song for help with animal pathology and experiments; E. Mesman, J. Bulthuis, K. de Goeij and M. Tjin-a-Koeng for histotechnical analysis; and L. Rijswijk, F. van der Ahé, H. Grimminck and L. Tolkamp for help with oncogenicity experiments and animal husbandry. We thank E. Koh and G. Daley for providing pEYK libraries; R. Beauchamp for RIE cells; K. Berns, B. Burgering, E. Danen, R. van der Kammen, R. Kortlever, S. Lens, F. Scheeren, S. Tait, E. de Vries and A. Werner for various reagents; R. Bernards, J. Borst, J. Collard, J. Hilken, W. Mooi, E. Roos, A. Sonnenberg, as well as the members of our laboratory and division, for helpful discussions; and M. van Lohuizen and A. Berns for suggestions and critical reading of the manuscript. D.S.P. and T.L. were supported by the Netherlands Organization for Scientific Research (NWO).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.S.P. (d.peepe@nki.nl).

Argos inhibits epidermal growth factor receptor signalling by ligand sequestration

Daryl E. Klein^{1*}, Valerie M. Nappi^{1*}, Gregory T. Reeves², Stanislav Y. Shvartsman² & Mark A. Lemmon¹

¹Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, 809C Stellar-Chance Laboratories, 422 Curie Boulevard, Philadelphia, Pennsylvania 19104-6059, USA

²Department of Chemical Engineering and the Lewis-Sigler Institute for Integrative Genomics, Princeton University, Carl Icahn Laboratory, Washington Road, Princeton, New Jersey 08544, USA

* These authors contributed equally to this work

The epidermal growth factor receptor (EGFR) has critical functions in development and in many human cancers^{1–3}. During development, the spatial extent of EGFR signalling is regulated by feedback loops comprising both well-understood activators and less well-characterized inhibitors^{3,4}. In *Drosophila melanogaster* the secreted protein Argos functions as the only known extracellular inhibitor of EGFR⁵, with clearly identified roles in multiple stages of development³. Argos is only expressed when the *Drosophila* EGFR (DER) is activated at high levels⁶, and downregulates further DER signalling. Although there is ample genetic evidence that Argos inhibits DER activation, the biochemical mechanism has not been established. Here we show that Argos inhibits DER signalling without interacting directly with the receptor, but instead by sequestering the DER-activating ligand Spitz. Argos binds tightly to the EGF motif of Spitz and forms a 1:1 (Spitz:Argos) complex that does not bind DER *in vitro* or at the cell surface. Our results provide an insight into the mechanism of Argos function, and suggest new strategies for EGFR inhibitor design.

Argos is a secreted *Drosophila* protein comprising 444 amino acids that has an atypical EGF-like motif at its carboxy terminus⁷. Argos was genetically identified as an inhibitor of DER signalling^{6–8}. Its expression is induced by DER activation^{5,6}, constituting a negative feedback mechanism that is involved in more than ten distinct DER-dependent processes during *Drosophila* development³. Several reports have argued that Argos interacts directly with DER to inhibit its signalling^{5,9,10}, leading us to hypothesize that Argos stabilizes an autoinhibited receptor configuration such as that revealed in our recent crystallographic studies of human EGFR¹¹.

To investigate how Argos inhibits DER signalling, we purified recombinant secreted Argos, Spitz (a TGF α -like DER-activating ligand³) and the DER2 extracellular region (sDER2) from trans-

fected *Drosophila Schneider-2* (S2) cells. Contrary to our expectations, sDER2 completely failed to interact with immobilized Argos in surface plasmon resonance (SPR) studies (Fig. 1a). Unexpectedly, Spitz bound to the same immobilized Argos with high affinity (Fig. 1a). These observations were not artefacts of Argos immobilization, as soluble Argos also bound strongly to immobilized Spitz ($K_d = 20 \pm 3$ nM) and did not bind to immobilized sDER2 (Fig. 1b). As shown in Fig. 1b and 1c, immobilized Spitz binds ~tenfold more strongly to Argos than to sDER2 (K_d for Spitz binding by sDER2 was 157 ± 37 nM, compared with 132 nM for EGF binding by human sEGFR¹¹). Furthermore analytical ultracentrifugation experiments showed that Argos and Spitz form a (monomeric) 1:1 complex in solution (Fig. 1d), and that Spitz induces the expected sDER2 dimerization (Fig. 1e). Thus, SPR and analytical ultracentrifugation experiments demonstrate that Argos binds directly to the activating ligand Spitz (and not to DER itself), in direct contrast to what was previously suggested from less quantitative studies^{5,9,10}. As shown in Supplementary Fig. S1, we identified the Argos binding site as the EGF-like domain of Spitz (residues 78–141), and found that Spitz binds to the C-terminal 220 amino acids of Argos, consistent with the ability of this Argos fragment to rescue loss-of-function alleles *in vivo*¹².

Because Argos and DER both bind the EGF-like domain of Spitz, we hypothesized that they may compete with one another for the same binding site on Spitz. Argos could thus 'sequester' Spitz away from DER by blocking the receptor-binding site on the activating ligand. This model has precedents in insulin-like growth factor (IGF)-binding proteins¹³ and bone morphogenetic protein (BMP) antagonists¹⁴, which both sequester their target ligands in this way. To test this hypothesis, we analysed the effect of added Argos on Spitz binding to immobilized sDER2 (Fig. 2a). On its own, Spitz (at 250 nM) bound immobilized sDER2 to give an SPR binding signal of ~600 response units (RUs). The addition of increasing amounts of Argos (to a fixed Spitz concentration) progressively reduced this signal until it reached zero, when Argos was present in slight excess (≥ 1.1 -fold). Because binding of either Spitz or Argos to immobilized sDER2 would contribute to the SPR signal, these results can only be explained if Argos binds directly to Spitz in solution (and not at all to immobilized sDER2), and thus blocks Spitz binding to the receptor. The finding that sDER2 binding is abolished when Argos and Spitz are approximately equimolar (performed at 12 times K_d for the Argos–Spitz interaction) is consistent with the 1:1 (Argos:Spitz) stoichiometry measured in Fig. 1d. We therefore conclude that Argos can act as a ligand 'sink' that efficiently sequesters Spitz in an inactive 1:1 complex and prevents Spitz from binding its receptor.

We next investigated whether Argos can act as a ligand sink in a cellular context. Addition of Spitz at 10 nM, 50 nM or 100 nM induces robust tyrosine autophosphorylation of DER2 expressed in S2 cells (Fig. 2b), and this is progressively inhibited by increasing amounts of Argos. As expected if Argos inhibits DER signalling by sequestering Spitz into an inactive complex, when more Spitz is present in Fig. 2b, more Argos is required to prevent it all from receptor binding. The requirement for a larger excess of Argos over Spitz (~fivefold) in this experiment compared with Fig. 2a suggests that Spitz binds more strongly to cell surface DER2 than to immobilized sDER2 (this is known to be true in the human EGF system¹⁵).

To visualize Spitz and Argos at the cell surface, we labelled recombinant proteins with Alexa Fluor-488 and -633 respectively, using the same chemistry as employed for SPR immobilization. As shown in Fig. 2c, fluorescently labelled Spitz-488 was seen clearly at the plasma membrane (and in compartments likely to represent endosomes) in ~40% of living D2f cells (approximately 50% of D2f cells expressed DER2 when analysed by FACS), but was not seen in parental S2 cells (that do not express DER). In contrast, fluorescent Argos-633 showed no DER-dependent binding to the cell surface.

Argos-633 was instead found at the surface of all S2 and D2f cells, regardless of DER expression, in a punctate pattern (see below). Thus, Argos binds neither to the isolated DER2 extracellular region (Fig. 1) nor to intact DER2 present at the cell surface (Fig. 2c). In an experiment analogous to the *in vitro* titration shown in Fig. 2a, we investigated whether Argos prevented Spitz-488 binding to the surface of DER2-expressing S2 cells. As shown in Fig. 2d, binding of Spitz-488 to DER2 at the cell surface was greatly reduced if it was added together with a fivefold excess of Argos—as expected if Argos functions as a ligand sink.

To gain additional insight into the mechanism of Argos action at the cell surface, we performed order-of-addition experiments. Excess Argos (tenfold) completely prevented both DER activation (Fig. 3a) and cell surface Spitz-488 binding (Fig. 3b) if it was added before Spitz (Aos → Spi), or as an Argos/Spitz mixture. However, if cells were first exposed to Spitz-488 alone, washed, and then treated with

excess Argos, DER activation was detected (Fig. 3a, Spi → Aos) and Spitz-488 remained bound to the cell surface (Fig. 3b, Spi → Aos). Thus, although Argos binds efficiently to free Spitz in solution and prevents it from binding the receptor, Argos did not seem to dissociate Spitz from pre-formed Spitz-DER complexes at the cell surface (on the time scale of these experiments). Under the conditions used here (with incubations on ice), activated DER would not be clustered into coated pits¹⁶, so would not be physically occluded from access to Argos. Experiments with human EGF using fluorescence-recovery-after-photobleaching¹⁷ indicate that ligand bound to activated (dimeric) EGFR exchanges very slowly, even before aggregation into coated pits. A similar slow exchange of DER-bound Spitz probably explains the lack of reversibility seen in our experiments. In longer time-course experiments (over 30 min at 25 °C), cell surface-associated Spitz-488 was internalized rather than being dissociated by excess Argos in the medium (data not

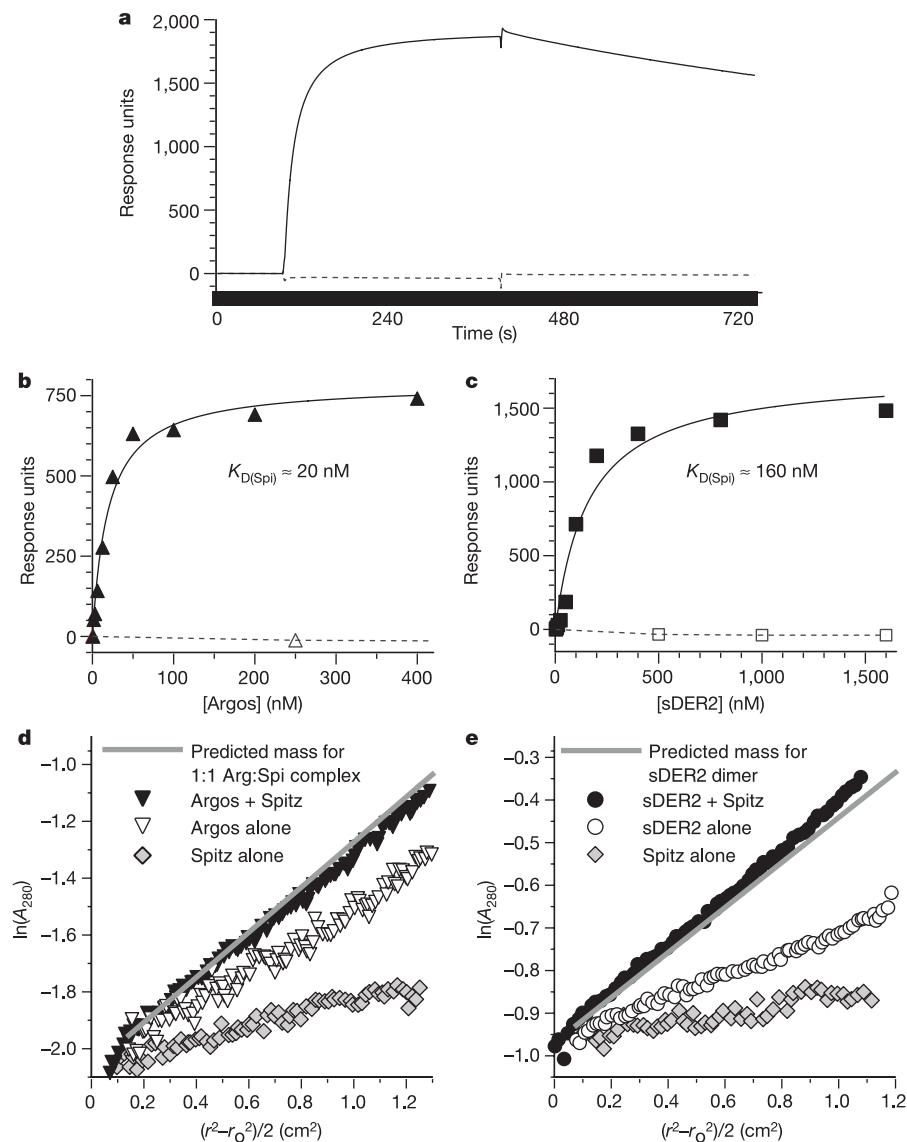


Figure 1 Quantitative analysis of Spitz/Argos interactions. **a**, SPR sensorgrams show that Spitz (at 1 μ M, solid line), but not sDER2 (at 1 μ M, dotted line) binds immobilized Argos. **b**, Representative binding curves for interaction of soluble Argos with immobilized Spitz (filled triangles) and immobilized sDER2 (open triangles). **c**, sDER2 binds immobilized Spitz (filled squares) but not immobilized Argos (open squares). **d**, In analytical ultracentrifugation, Argos (open triangles) sediments at 48.8 ± 4.3 kDa (47.6 kDa expected without glycosylation), whereas Spitz (diamonds) sediments as a

16.2 ± 1.7 kDa species (11.8 kDa expected without glycosylation). Equimolar Argos/Spitz mixtures fit best to a single ~ 59 kDa species (filled triangles). The grey straight line represents expected results for a 1:1 Argos:Spitz complex. **e**, sDER2 (open circles) sediments as a 101.8 ± 0.6 kDa species, which approximately doubles upon adding a 1.2-fold excess of Spitz (filled circles); Spitz alone (diamonds) sediments as in **d**. The grey line represents expected results for an sDER2 dimer.

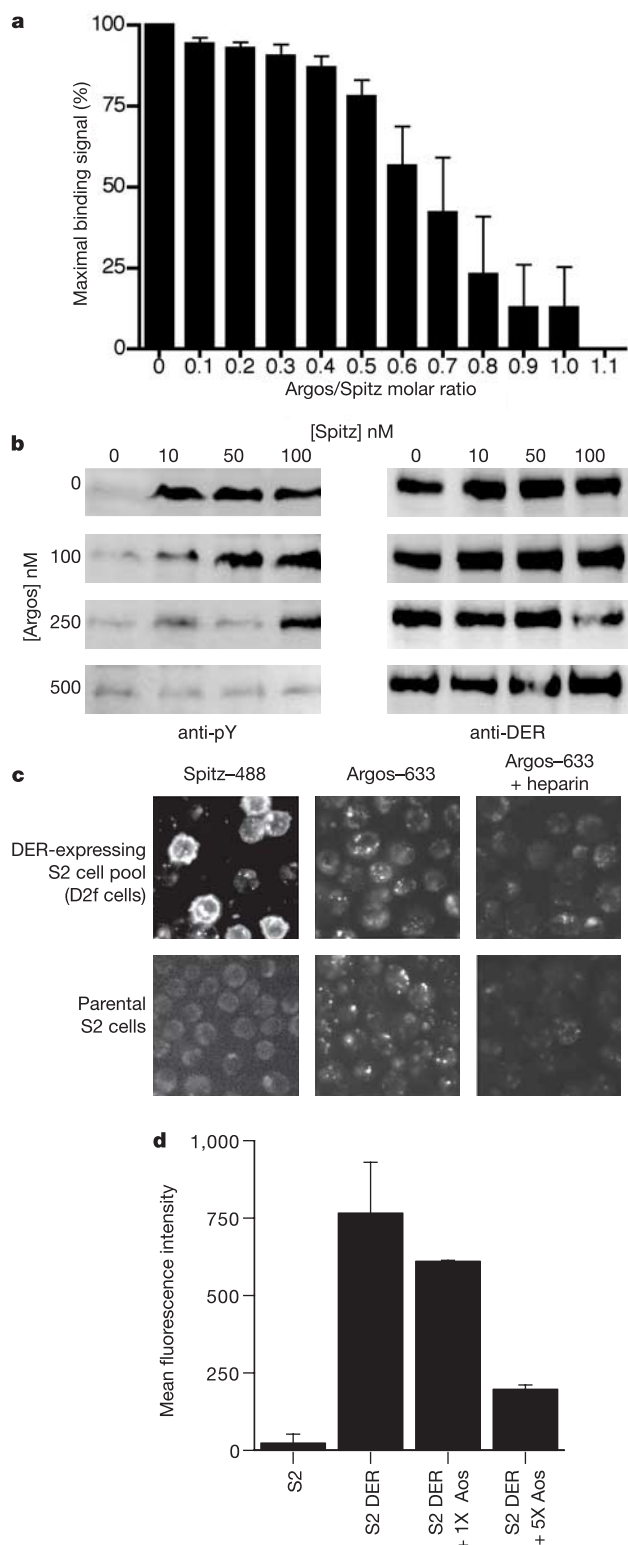


Figure 2 Argos sequesters Spitz away from DER. **a**, Binding to immobilized sDER2 was measured for samples containing 250 nM Spitz and the indicated mole fraction of Argos. Binding is abolished when Argos is in excess. Mean $\pm$ s.d. ($n \geq 3$) is shown. **b**, Western blot showing robust Spitz-induced tyrosine phosphorylation of DER2 expressed in S2 cells and the effect of adding different concentrations of Argos, using an anti-phosphotyrosine antibody (anti-pY, left panels). Western blot showing DER2 protein levels using an anti-DER2 antibody (right panels). **c**, Spitz-488 binds to the $\sim 40\%$ of living D2f cells that express DER, but not to parental S2 cells. Argos-633 gives an identical punctate staining pattern on both parental and DER-expressing S2 cells, which is largely abolished by $10 \mu\text{g ml}^{-1}$ low-molecular-weight heparin. **d**, A fivefold excess of Argos prevents cell-surface labelling by Spitz-488. Mean fluorescence intensity is defined in the Methods.

shown). We therefore conclude that Argos efficiently blocks the initiation of DER activation by Spitz (by acting as a sink for soluble ligand), but does not reverse DER activation once underway. The activated receptor is internalized before bound Spitz can dissociate and interact with Argos. As receptor-bound Spitz seems to be less susceptible to Argos inhibition than free Spitz, this finding also lends further support to the argument that Argos does not exert its effects through DER binding.

The punctate distribution of Argos-633 on the surface of all S2 and D2f cells (Fig. 2c) resembles that seen for Noggin, a BMP antagonist reported to bind heparan sulphate proteoglycans (HSPGs) at the cell surface¹⁸. Like Noggin¹⁸, we found that Argos could be dissociated from the cell surface by adding excess low-molecular-weight (5 kDa) heparin (Fig. 2c), and also that it could bind efficiently to heparin-sepharose beads (Fig. 3c). However, low-molecular-weight heparin did not mimic all properties of the cell surface Argos binding site. Excess heparin did not prevent Argos from blocking DER activation by Spitz in Fig. 3a, and Argos-Spitz complexes seemed to bind heparin-sepharose beads (Fig. 3c). By contrast, the experiments in Figs 2d and 3b argue that cell surface associated Argos does not bind Spitz. If it did, Spitz-488 should bind to parental S2 cells in an Argos-dependent manner—a possibility that our experiments in Fig. 3b rule out. Heparin frequently fails to mimic all aspects of physiological interactions with cell surface HSPGs, which may involve specific oligosaccharide sequences not found in heparin¹⁹ and/or may require contacts with the HSPG protein core²⁰. Such specific interactions may be responsible for preventing cell surface Argos from binding Spitz, and their further characterization awaits identification of the cell surface binding partner for Argos.

Our *in vitro* and cellular findings refute previous suggestions that Argos is an antagonistic ligand (or inverse agonist) of DER^{5,9,10}. The only evidence for direct Argos-DER interactions came from a chemical crosslinking study⁹ and weak SPR signals that may represent non-specific background binding⁹. To our knowledge, Argos-Spitz interactions have not previously been analysed. As a direct DER antagonist (or inverse agonist), Argos would be unprecedented in receptor tyrosine kinase signalling. In fact, across all receptors with a single membrane span, we know of only one inhibitor that functions in this way; the interleukin-1 receptor antagonist (IL-1Ra)²¹. Inhibition of signalling by binding to (and sequestering) stimulatory ligands is much more common, and also seems to be how Argos functions. This mechanism is used by soluble decoy receptors such as osteoprotegerin²², by the IGF-binding proteins¹³ and by BMP antagonists such as Noggin, Chordin and DAN family proteins¹⁴. Interestingly, just as Spitz and its antagonist (Argos) both contain critical EGF-like motifs, so are BMPs and their antagonists (for example, Noggin) structurally related, by sharing a cystine knot topology¹⁴.

Our finding that Argos functions as a Spitz-induced Spitz antagonist suggests a new model for its mechanism of action *in vivo*. Rather than acting as a direct long range DER inhibitor as previously hypothesized^{3,5,7}, the fact that Argos sequesters Spitz and is induced by high levels of Spitz (or other DER agonists) suggests that the long range effects of Argos may be mediated by its influence on Spitz distribution. In other words, Argos may act locally to restrict further Spitz signalling and/or to limit delivery of active Spitz to distant cells. Computational modelling of DER/Spitz/Argos interactions in tissues^{23,24}, based on the findings reported here, is required to evaluate this possibility. Our preliminary modelling results with *Drosophila* ventral ectoderm patterning suggest that efficient Spitz sequestration by Argos is key for explaining the existing data and for providing a robust feedback mechanism that modulates the secreted Spitz gradient. In our models, Argos action over only a short range can effectively limit the longer range of Spitz signalling.

Future studies are required to determine whether the other

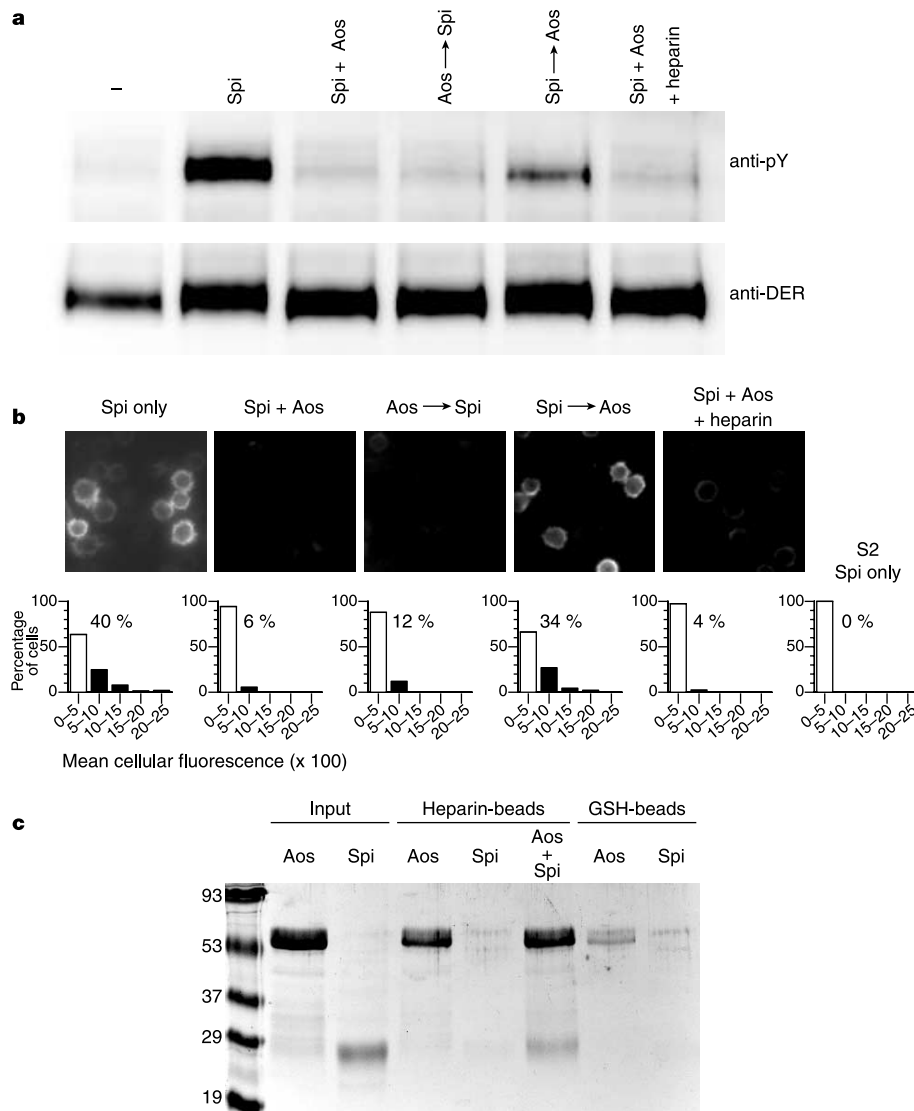


Figure 3 Argos does not displace Spitz pre-bound to cell surface DER. **a**, Western blot showing DER phosphorylation in serum-starved D2f cells left untreated (—), treated with 30 nM Spitz alone on ice (Spi) or treated with 30 nM Spitz plus 300 nM Argos in the order indicated. **b**, Parallel experiments analysing the effect of adding 1 μM unlabelled Argos on Spitz-488 binding (at 100 nM) to living D2f cells. Histograms describe cellular distribution of fluorescence intensity for each experiment (see Methods), plus a control experiment in which Spitz-488 was added to parental cells ('S2 Spi only'). The percentage of cells with

fluorescence intensity >500 (filled bars) is reported above each graph. Spitz-488 alone, Spi; pre-mixed Spitz and Argos, Spi + Aos; Aos → Spi and Spi → Aos are defined in the text; addition of premixed Argos, Spitz and heparin, Spi + Aos + heparin (30 μg ml⁻¹ in **a**, 20 μg ml⁻¹ in **b**). **c**, Coomassie-stained SDS-PAGE gel shows Argos (50 μg ml⁻¹ in input) but not Spitz (4 μg ml⁻¹ in input) sedimenting with excess heparin-sepharose beads (washed with 150 mM NaCl buffer). Both Argos and Spitz sediment from an Argos/Spitz mixture, indicating that heparin-bound Argos retains Spitz-binding capacity.

EGFR-activating ligands in *Drosophila*³ (Gurken, Keren and Vein) are also affected by Argos, and whether a mammalian Argos equivalent exists. Finally, our demonstration that Argos efficiently sequesters an activating ligand may provide the basis for developing novel related agents that can neutralize EGFR ligands when their overexpression contributes to human cancers^{2,25}. □

Methods

Production of Argos, Spitz and sDER2

A hexahistidine tag followed by a stop codon was introduced by polymerase chain reaction (PCR) after the C terminus of Argos (residue 444), after Pro 128 of Spitz and after Lys 861 of the DER2 extracellular region. Coding regions were subcloned into pMT/V5-HisA (Invitrogen) for S2 cell secretion (using the native signal sequences). The Spitz EGF domain (amino acids 67–128) and C-terminal Argos fragment (amino acids 224–444) used the BiP signal sequence. Plasmids were co-transfected with pCo-Hygro into S2 cells, and expressing pools were selected with hygromycin. Resistant cells were grown to 4 × 10⁹ cells per ml in glutamine-supplemented DES serum-free expression medium (Invitrogen), and induced for 3–4 days with 0.5 mM Cu₂SO₄. Medium was collected, diafiltered against 3–5 volumes of 150 mM NaCl and 25 mM HEPES at pH 8, and protein was purified as

described²⁶. As previously reported²⁷, secreted Spitz has an anomalously high apparent molecular weight (~28 kDa) in SDS-polyacrylamide gel electrophoresis (PAGE) (Fig. 3c), as frequently seen for small glycoproteins. Mass spectrometry (not shown) and analytical ultracentrifugation (Fig. 1d) confirmed that it was ~16 kDa. The identity of purified Argos was confirmed by amino-terminal amino acid sequencing. All *in vitro* measurements were performed in 10 mM Hepes buffer at pH 7.4, containing 150 mM NaCl, 3 mM EDTA and 0.005% Surfactant P20 at 25 °C.

SPR analysis

Proteins were immobilized on CM5 sensorchips by amine-coupling as described²⁶, with pre-concentration in acetate buffer at pH 5 (Argos and Spitz), pH 5.5 (sDER2) or pH 3.0 (Spitz EGF motif). The signal from immobilization was 5,500 RUs (Spitz), 8,000 RUs (Argos), 14,000 RUs (sDER2) and 700 RUs (Spitz EGF motif). Binding curves were generated and analysed as described²⁶, using 50 μl injections (at 10 μl min⁻¹) for sDER2/Spitz interactions, and 240 μl injections for Argos/Spitz binding (to ensure that steady state was reached). Binding surfaces were regenerated between points using 10 μl of 1 M NaCl in 10 mM glycine, pH 3.0 (for Argos/Spitz binding) or in 10 mM acetate, pH 4.5 (sDER2/Spitz binding). Controls established that regeneration did not impair binding capacity (or affinity characteristics) of each sensorchip surface in a given experiment. K_d is quoted as the mean ± s.d. for at least three experiments.

Analytical ultracentrifugation

Sedimentation equilibrium experiments were performed at 25 °C with 5 µM protein in a Beckman XL-A centrifuge, and analysed exactly as described²⁶. Rotor speeds were 8,000 rpm for Fig. 1d and 5,000 rpm for Fig. 1e. Data are plotted in Fig. 1 as the natural logarithm of the 280 nm absorbance ($\ln(A_{280})$) against a function of the square of the radius ($(r^2 - r_0^2)/2$). For a single species, this plot gives a straight line with gradient proportional to molecular mass. Molecular masses are quoted as the means $\pm$ s.d. of at least three independent determinations.

Analysis of DER activation

DER2-expressing S2 cells (D2f) were kindly provided by B.-Z. Shilo²⁷. Cells were serum-starved overnight. DER expression was induced for 3 h with 60 µM Cu₂SO₄, and the receptor was activated with the indicated Spitz concentrations (in Figs 2b and 3a) on ice for 10 min, in 25 mM Tris-HCl at pH 7.4, 120 mM NaCl, 5 mM KCl, 1.2 mM MgCl₂, 0.1% (w/v) bovine serum albumin and 1 mM dextrose. Cells were lysed in radioimmunoprecipitation assay (RIPA) buffer containing phosphatase inhibitors. Clarified RIPA-lysates were analysed by western blotting with anti-phosphotyrosine antibodies (pY20; Santa Cruz) and (for normalization) a rabbit polyclonal antibody against the DER C terminus (a gift from J. Duffy), followed by relevant HRP-conjugated secondary antibodies for enhanced chemiluminescence detection.

Analysis of cell surface binding by fluorescently-labelled proteins

Purified Spitz and Argos were labelled with Alexa Fluor-488 and -633 kits respectively (Molecular Probes), to give Spitz-488 and Argos-633. Labelled (or unlabelled) proteins were added at 100 nM to DER2-expressing D2f cells (serum-starved overnight) in PBS/1% BSA. Incubations were performed at room temperature (Fig. 2c, d) or on ice (Fig. 3b). Living cells were washed three times with PBS/1% BSA, and inspected by fluorescence microscopy with appropriate filter-sets at room temperature. The mean pixel intensity was calculated for each cell in the field (values ranged from 0 to 3,000), corrected for background fluorescence, and this value was averaged over all cells (>100) to give the 'fluorescence intensity' (in arbitrary units) for that experiment. The mean of this value ($\pm$ s.d.) was plotted (Fig. 2d). For order-of-addition experiments (Fig. 3b), cells were treated (on ice) for 1 min with the first ligand (or mixture), rinsed with cold PBS/1% BSA, and then incubated (on ice) with the second ligand (or mixture). Living cells were then inspected by fluorescence microscopy at room temperature. The mean pixel intensity was calculated for each cell (range from 0 to 3,000), and its distribution for each experiment was plotted (Fig. 3b). To investigate HSPG-dependent binding of Argos to the cell surface, low molecular weight (5 kDa) heparin was added to cells alongside Argos-633 (where indicated) at a final concentration of 10 µg ml⁻¹.

Received 18 May; accepted 7 July 2004; doi:10.1038/nature02840.

- Jorissen, R. N. *et al.* Epidermal growth factor receptor: mechanisms of activation and signalling. *Exp. Cell Res.* **284**, 31–53 (2003).
- Yarden, Y. & Slivkowski, M. X. Untangling the ErbB signalling network. *Nature Rev. Mol. Cell Biol.* **2**, 127–137 (2001).
- Shilo, B. Z. Signaling by the *Drosophila* epidermal growth factor receptor pathway during development. *Exp. Cell Res.* **284**, 140–149 (2003).
- Wiley, H. S., Shvartsman, S. Y. & Lauffenburger, D. A. Computational modeling of the EGF-receptor system: a paradigm for systems biology. *Trends Cell Biol.* **12**, 43–50 (2003).
- Schweitzer, R., Howes, R., Smith, R., Shilo, B. Z. & Freeman, M. Inhibition of *Drosophila* EGF receptor activation by the secreted protein Argos. *Nature* **376**, 699–702 (1995).
- Golembo, M., Schweitzer, R., Freeman, M. & Shilo, B. Z. Argos transcription is induced by the *Drosophila* EGF receptor pathway to form an inhibitory feedback loop. *Development* **122**, 223–230 (1996).
- Freeman, M., Klambt, C., Goodman, C. S. & Rubin, G. M. The *argos* gene encodes a diffusible factor that regulates cell fate decisions in the *Drosophila* eye. *Cell* **69**, 963–975 (1992).
- Wasserman, J. D. & Freeman, M. An autoregulatory cascade of EGF receptor signaling patterns the *Drosophila* egg. *Cell* **95**, 355–364 (1998).
- Jin, M. H., Sawamoto, K., Ito, M. & Okano, H. The interaction between the *Drosophila* secreted protein argos and the epidermal growth factor receptor inhibits dimerization of the receptor and binding of secreted spitz to the receptor. *Mol. Cell Biol.* **20**, 2098–2107 (2000).
- Vinos, J. & Freeman, M. Evidence that Argos is an antagonistic ligand of the EGF receptor. *Oncogene* **19**, 3560–3562 (2000).
- Ferguson, K. M. *et al.* EGF activates its receptor by removing interactions that autoinhibit ectodomain dimerization. *Mol. Cell* **11**, 507–517 (2003).
- Howes, R., Wasserman, J. D. & Freeman, M. *In vivo* analysis of Argos structure-function. Sequence requirements for inhibition of the *Drosophila* epidermal growth factor receptor. *J. Biol. Chem.* **273**, 4275–4281 (1998).
- Duan, C. Beyond carrier proteins: Specifying the cellular responses to IGF signals: roles of IGF-binding proteins. *J. Endocrinol.* **175**, 41–54 (2002).
- Groppe, J. *et al.* Structural basis of BMP signalling inhibition by the cystine knot protein Noggin. *Nature* **420**, 636–642 (2002).
- Mattoon, D., Klein, P., Lemmon, M. A., Lax, I. & Schlessinger, J. The tethered configuration of the EGF receptor extracellular domain exerts only a limited control of receptor function. *Proc. Natl Acad. Sci. USA* **101**, 923–928 (2004).
- Schlessinger, J., Schecter, Y., Willingham, M. C. & Pastan, I. Direct visualization of binding, aggregation, and internalization of insulin and epidermal growth factor on living fibroblastic cells. *Proc. Natl Acad. Sci. USA* **75**, 2659–2663 (1978).
- Verveer, P. J., Wouters, F. S., Reynolds, A. R. & Bastiaens, P. I. Quantitative imaging of lateral ErbB1 receptor signal propagation in the plasma membrane. *Science* **290**, 1567–1570 (2000).
- Paine-Saunders, S., Viviano, B. L., Economides, A. N. & Saunders, S. Heparan sulfate proteoglycans retain Noggin at the cell surface: a potential mechanism for shaping bone morphogenetic protein gradients. *J. Biol. Chem.* **277**, 2089–2096 (2002).

- Powell, A. K., Yates, E. A., Fernig, D. G. & Turnbull, J. E. Interactions of heparin/heparan sulfate with proteins: appraisal of structural factors and experimental approaches. *Glycobiology* **14**, 17R–30R (2004).
- Kramer, K. L. & Yost, H. J. Heparan sulfate core proteins in cell–cell signaling. *Annu. Rev. Genet.* **37**, 461–484 (2003).
- Arend, W. P., Malyak, M., Guthridge, C. J. & Gabay, C. Interleukin-1 receptor antagonist: role in biology. *Annu. Rev. Immunol.* **16**, 27–55 (1998).
- Mantovani, A., Locati, M., Vecchi, A., Sozzani, S. & Allavena, P. Decoy receptors: a strategy to regulate inflammatory cytokines and chemokines. *Trends Immunol.* **22**, 328–336 (2001).
- Barkai, N. & Shilo, B. Z. Modeling pattern formation: counting to two in the *Drosophila* egg. *Curr. Biol.* **12**, R493–R495 (2002).
- Shvartsman, S. Y., Muratov, C. B. & Lauffenburger, D. A. Modeling and computational analysis of EGF receptor-mediated cell communication in *Drosophila* oogenesis. *Development* **129**, 2577–2589 (2002).
- Salomon, D. S., Brandt, R., Ciardiello, F. & Normanno, N. Epidermal growth factor-related peptides and their receptors in human malignancies. *Crit. Rev. Oncol. Hematol.* **19**, 183–232 (1995).
- Ferguson, K. M., Darling, P. J., Mohan, M. J., Macatee, T. L. & Lemmon, M. A. Extracellular domains drive homo- but not hetero-dimerization of erbB receptors. *EMBO J.* **19**, 4632–4643 (2000).
- Schweitzer, R., Shahrabany, M., Seger, R. & Shilo, B. Z. Secreted Spitz triggers the DER signaling pathway and is a limiting component in embryonic ventral ectoderm determination. *Genes Dev.* **9**, 1518–1529 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Duffy, K. Ferguson, P. Carroll, G. Van Duynne and members of the Lemmon and Shvartsman laboratories for valuable discussions; T. Schüpbach and N. Perrimon for providing cDNAs. This work was supported by grants from the NIH (to M.A.L.) and NSF (S.Y.S.), by NIH training grant support (to D.E.K. and V.M.N.), and by an NSF graduate research fellowship (G.T.R.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.A.L. (mlemmon@mail.med.upenn.edu).

Separase-mediated cleavage of cohesin at interphase is required for DNA repair

Koji Nagao^{1,3}, Yoh Adachi² & Mitsuhiro Yanagida^{1,2,3}

¹Department of Biophysics, Graduate School of Science and ²Department of Gene Mechanisms, Graduate School of Biostudies, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

³Initial Research Project, Okinawa Institute of Science and Technology, Okinawa 904-2234, Japan

Sister chromatids are held together by cohesins¹. At anaphase, separase is activated by degradation of its inhibitory partner, securin^{2,3}. Separase then cleaves cohesins^{4–6}, thus allowing sister chromatid separation. Fission yeast securin (Cut2) has destruction boxes and a separase (Cut1) interaction site in the amino and carboxyl terminus, respectively^{7,8}. Here we show that securin is essential for separase stability and also for proper repair of DNA damaged by ultraviolet, X-ray and γ -ray irradiation. The *cut2*^{EA2} mutant is defective in the repair of ultraviolet damage lesions, although the DNA damage checkpoint is activated normally. In double mutant analysis of ultraviolet sensitivity, checkpoint kinase *chk1* (ref. 9) and excision repair *rad13* (ref. 10) mutants were additive with *cut2*^{EA2}, whereas recombination repair *rhp51* (ref. 11) and cohesin subunit *rad21* (ref. 12) mutants were not. Cohesin was hyper-modified on ultraviolet irradiation in a Rad3 kinase-dependent way¹³. Experiments using either mutant cohesin that cannot be cleaved by separase or a protease-dead separase provide evidence that this DNA repair function of securin–separase acts through the cleavage of cohesin. We propose that

the securin-separase complex might aid DNA repair by removing local cohesin in interphase cells.

We introduced several substitutions into the central domain of securin/Cut2 by homologous recombination in *Schizosaccharomyces pombe*. One mutant designated EA2 (amino acids 127–129, DIE → AIA, Fig. 1a) turned out to be temperature sensitive (Supplementary Fig. 1a). The temperature sensitive phenotype of

cut2^{EA2} was completely suppressed by a plasmid that carried the wild-type *cut2*⁺ gene (data not shown). Severe anaphase defects were found in *cut2*^{EA2} at 36 °C (Supplementary Fig. 1b). Sister chromatids did not separate, and although spindle elongation was greatly diminished, cytokinesis followed, resulting in a *cut* phenotype. Although the *cut2*^{EA2} phenotype resembled that of indestructible *cut2* (ref. 2), Cut2^{EA2} protein was rapidly degraded in anaphase

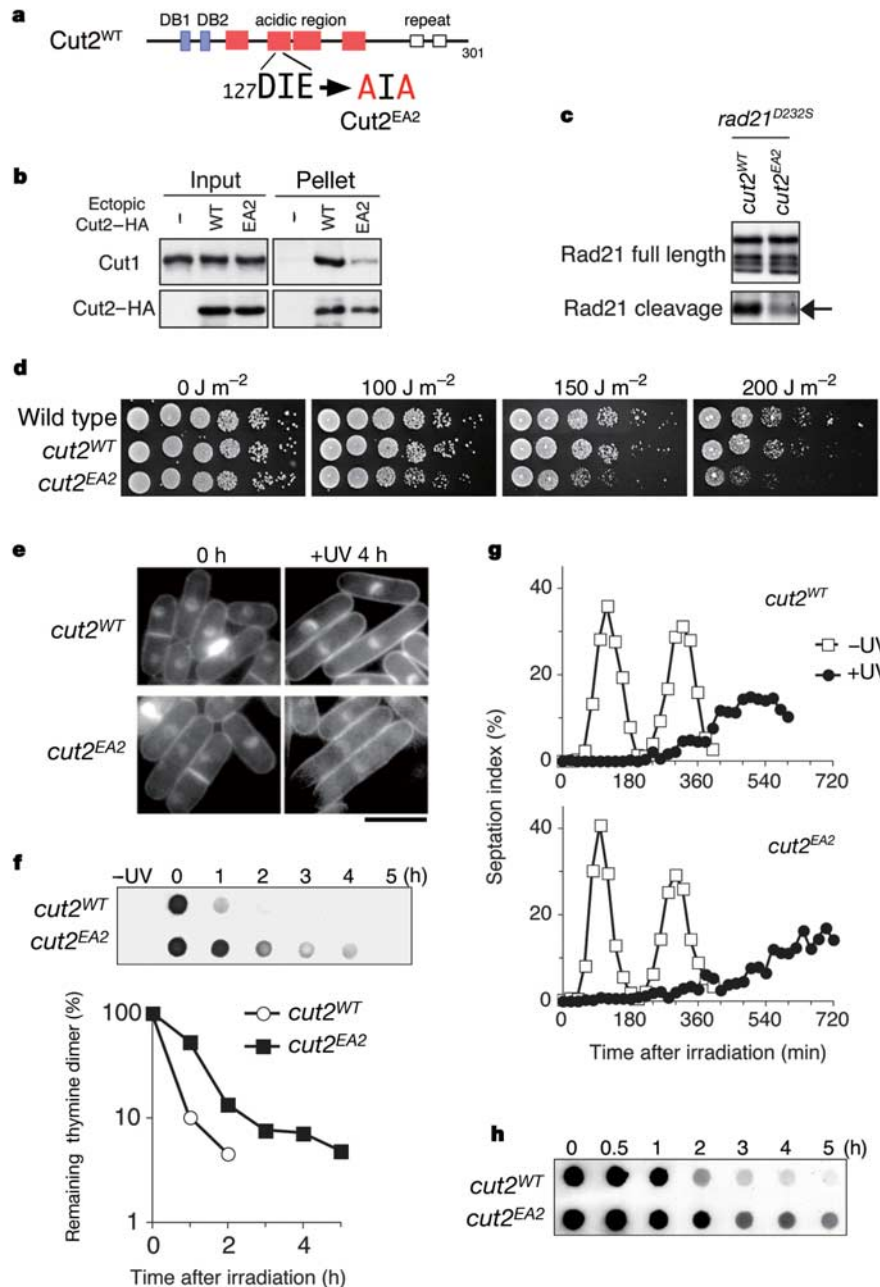


Figure 1 Securin mutant *cut2*^{EA2} is defective in DNA damage repair in checkpoint-arrested cells. **a**, Chromosomally-integrated mutation *cut2*^{EA2}. DB1, DB2: destruction box sequences. **b**, Wild-type cells carrying pREP81-*cut2*^{WT}-HA or pREP81-*cut2*^{EA2}-HA were cultured at 33 °C for 11 h in the absence of thiamine (promoter on). Resulting cell extracts were immunoprecipitated using anti-HA antibodies. **c**, The Rad21 cleavage product of *cut2*^{WT} and *cut2*^{EA2} at 26 °C in a *rad21*^{D232S} background is indicated by an arrow. **d**, Serial dilutions of wild type (haploid 972), *cut2*^{WT} (integrated with the wild-type *cut2*⁺ gene construct) and *cut2*^{EA2} cells were spotted at 26 °C onto YPD plates, then irradiated with the dose indicated. **e**, Asynchronous cultures of *cut2*^{WT} and *cut2*^{EA2} were

exposed to ultraviolet irradiation (UV; 100 J m⁻²), and cells were stained by DAPI. Scale bar is 10 μm. **f**, The level of thymine dimers in genomic DNA from asynchronous cells after ultraviolet irradiation (100 J m⁻²) (top). Quantitative representation of the top panel (bottom). **g**, Wild type (*cut2*^{WT}) and mutant *cut2*^{EA2} were synchronized in early G2 by an elutriation, then irradiated (100 J m⁻²) at time 0. Cell cycle progression after ultraviolet irradiation or without irradiation was monitored by scoring the septation index. Control cells (–UV) showed two highly synchronous cell cycle advancements. **h**, The level of thymine dimer in genomic DNA from irradiated cells shown in **g** (+UV).

similar to wild type cells (Supplementary Fig. 1c); from this we conclude that DIE is not required for proteolysis of securin in anaphase.

The level of separase was diminished in the *cut2^{EA2}* mutant (Supplementary Fig. 2), suggesting that it might dissociate from Cut2^{EA2} and become susceptible to proteolysis. To examine this possibility, wild-type Cut2 and mutant Cut2^{EA2} tagged with haemagglutinin antigen (HA) were used to co-immunoprecipitate Cut1. As the level of co-immunoprecipitated Cut1 was greatly diminished in cells expressing Cut2^{EA2}-HA (Fig. 1b), the DIE sequence seems to be required for separase stability.

It is known in budding yeast that the cleaved cohesin product is especially unstable because of N-end rule dependent degradation¹⁴. We introduced an N-end rule resistant mutation (D232S) into the chromosomal *rad21* gene and examined the level of Rad21 cleavage product. In this background, the level of the Rad21 cleavage product (Fig. 1c, indicated by arrow) was much higher than in the wild-type background (see below). In *cut2^{EA2}* cells the level of Rad21^{D232S} cleavage product was lower than that in *cut2^{WT}* even at the permissive temperature (Fig. 1c), suggesting that separase activity was reduced in the *cut2^{EA2}* mutant. Thus one of the functions of securin seems to be as a chaperone for separase, consistent with previous results with human securin¹⁵.

The temperature sensitive *cut2^{EA2}* strain grew normally at 26°C. Therefore, we looked for a role of securin/Cut2 other than in anaphase and found that *cut2^{EA2}* was sensitive to ultraviolet irradiation (100–200 J m⁻²) at 26°C (Fig. 1d). The degree of ultraviolet sensitivity was not as strong as in *rad3/ATR* or *crb2*

mutants but was similar to *chk1* and *cds1* kinase mutants^{9,16,17}, all of which have been implicated in the DNA damage checkpoint. Tetrad analysis for *cut2^{EA2}* indicated that both the ultraviolet sensitivity and the temperature sensitive phenotypes were caused by the same EA2 mutation, and not an additional silent mutation (Supplementary Fig. 3).

Mutant *cut2^{EA2}* cells became elongated on irradiation at 26°C (Fig. 1e) and the cell division index (septation percentage) decreased, suggesting that the DNA damage checkpoint seemed to be normally activated. We found that the completion of the repair process in *cut2^{EA2}* was delayed compared with the wild type (Fig. 1f). The level of thymine dimers persisted after irradiation, suggesting that the DNA repair process was partly impaired, despite activation of the DNA damage checkpoint. To rigorously test whether the checkpoint arrest was fully activated and the repair process was delayed in the arrested cells, synchronous cultures were collected in the presence or absence of ultraviolet irradiation. Early G2 cells collected from wild-type and mutant *cut2^{EA2}* cultures were irradiated at time 0 and cultured at 26°C (Fig. 1g). After ultraviolet irradiation, cells were arrested for the period of two cell cycles in both cultures. The recovery of damaged DNA lesions (the level of thymine dimers) was rather slow in the *cut2^{EA2}* mutant culture in comparison with the wild type (Fig. 1h), establishing that the DNA damage checkpoint in *cut2^{EA2}* was intact whereas repair was impaired in the G2 phase.

By double mutant analysis in ultraviolet-irradiated cells, we found additive effects between *cut2^{EA2}* and the following mutants:

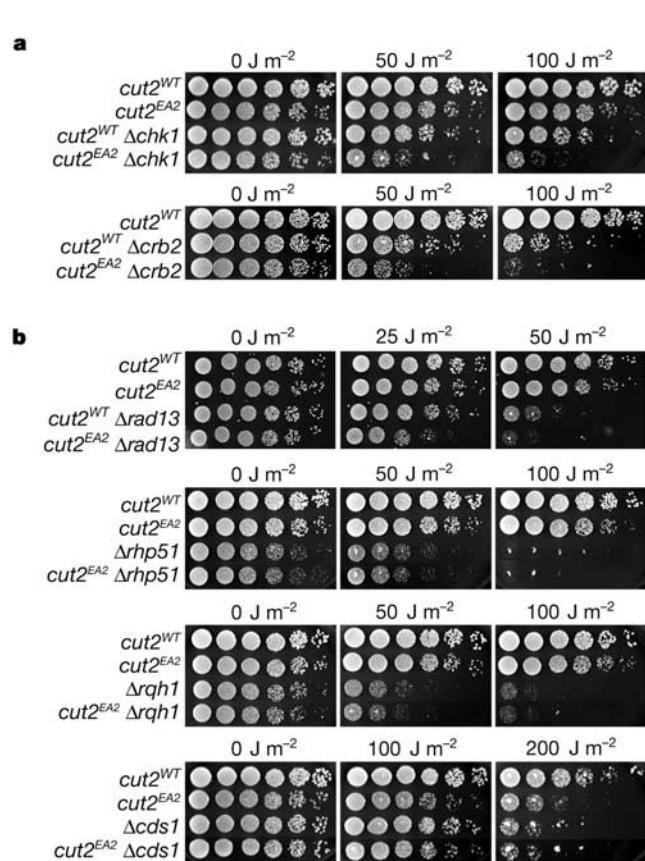


Figure 2 Ultraviolet sensitivity of *cut2^{EA2}* is additive with *Δchk1*, but not with *Δrhp51*. **a**, Additive hypersensitivity of double mutants of *cut2^{EA2}* with *Δchk1* or *Δcrb2* to ultraviolet irradiation at 26°C. **b**, Additive sensitivity to ultraviolet irradiation of double mutant *cut2^{EA2} Δrad13*. Double mutants of *cut2^{EA2}* with *Δrhp51*, *Δrqh1* or *Δcds1* showed no additive sensitivity to ultraviolet irradiation.

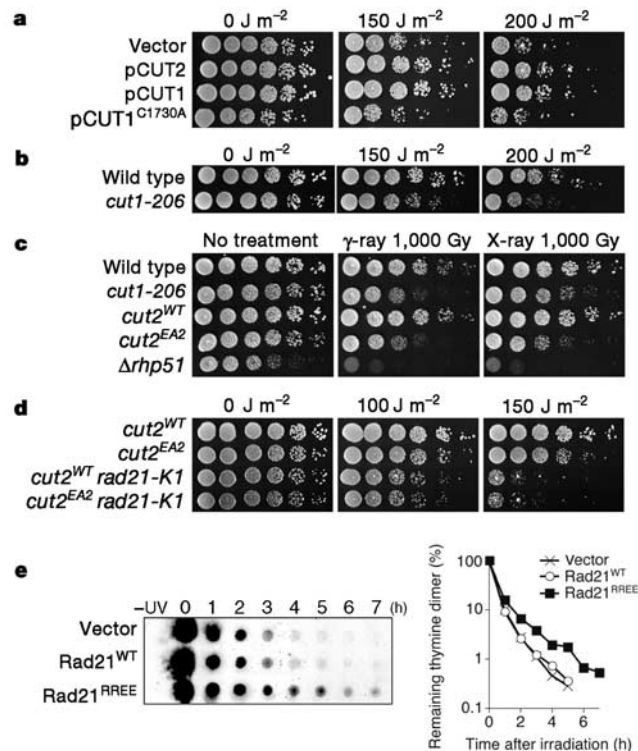


Figure 3 The DNA damage repair role of securin-separase acts through Rad21 cohesin. **a**, The ultraviolet sensitivity of *cut2^{EA2}* was suppressed by pCUT1 and pCUT2 but not by pCUT1^{C1730A} carrying the separase-dead gene. **b**, The separase mutant *cut1-206* was sensitive to ultraviolet irradiation (200 J m⁻²) at 26°C. **c**, The indicated strains were irradiated with γ-rays (1,000 Gy) or X-rays (1,000 Gy). **d**, Ultraviolet sensitivity of the double mutants *cut2^{EA2} rad21-K1* with three control strains. **e**, The level of thymine dimers in genomic DNA from ultraviolet (UV)-irradiated (100 J m⁻² at 26°C) wild-type cells moderately overexpressing wild-type Rad21 (Rad21^{WT}) or non-cleavable mutant Rad21^{REE} (Rad21^{REE}) (left). Quantitative estimation of the data shown on the left (right).

checkpoint kinase $\Delta chk1$ (ref. 9) or $chk1$ kinase dead (KD); $\Delta crb2$ (ref. 16), which is defective in damage checkpoint; and excision repair nuclease $\Delta rad13$ mutant (ref. 10) (Fig. 2a, b; and Supplementary Fig. 4a). Double mutants with $\Delta rgh1$ (ref. 18), $\Delta rhp51$ (ref. 11), $\Delta cds1$ (ref. 19) or $\Delta uvrde$ (ref. 20), however, did not show such an additive effect (Fig. 2b; and Supplementary Fig. 4b), implicating Cut2 in DNA double-strand break repair. The ultraviolet sensitivity of $cut2^{EA2}$ was thus a unique phenotype accompanied with the reduction of separase stability.

To identify genes that were related to securin-dependent DNA repair, multicopy plasmids that could suppress the temperature sensitive phenotype of $cut2^{EA2}$ were screened using a genomic DNA library of *S. pombe*. The majority (82 clones) carried the $cut2^+$ gene, whereas 36 clones contained the $cut1^+$ gene. Overexpression by pCUT1 may restore the level of unstable Cut1 in the $cut2^{EA2}$ mutant at 36°C. The ultraviolet sensitivity of $cut2^{EA2}$ at 26°C was rescued by pCUT1 but not by pCUT1^{C1730A} carrying the separase dead mutation²¹ (Fig. 3a), strongly suggesting that the damage repair required separase activity.

We examined whether any temperature sensitive $cut1$ alleles were ultraviolet sensitive at 26°C and identified $cut1-206$ (Fig. 3b). Although the degree of ultraviolet sensitivity was less than that in $cut2^{EA2}$, tetrad analysis indicated that the temperature sensitive and ultraviolet sensitivity phenotypes were caused by the same mutation (data not shown). Furthermore it was found that $cut1-206$ and $cut2^{EA2}$ were sensitive to X-ray and γ -ray irradiation (Fig. 3c). These

results suggest that both securin and separase are involved in DNA damage repair processes that include double-strand break repair.

To understand the relationship between the DNA damage repair role of securin–separase and the potential target cohesin, several experiments were performed. First, we examined the ultraviolet sensitivity of the double mutant $cut2^{EA2} rad21-K1$ (Fig. 3d). The ultraviolet sensitivity of $rad21-K1$ and the double mutant did not change, consistent with the hypothesis that securin and Rad21 (ref. 12) belong to the same repair pathway. Second, we tested whether wild-type cells expressing the non-cleavable mutant Rad21^{RREE} (ref. 22) were normal in the repair of thymine dimers produced by ultraviolet irradiation. As shown in Fig. 3e, the repair of thymine dimers was clearly slower than the control wild-type strains. These results strongly suggested that the cleavage of cohesin at the site of separase was required for the normal DNA damage repair.

Furthermore, we found that Rad21 was responsive to irradiation (Fig. 4a, left panel), producing a hyper-shifted upper band 15 min after irradiation (indicated by an asterisk). The anaphase cleaved product observed in asynchronously growing cells (indicated by an arrow) disappeared upon irradiation, probably due to the G2 cell cycle arrest mediated by the DNA damage checkpoint and the short protein half life. Two radiation-induced bands indicated by dots were accompanied by the appearance of the hyper-shifted upper Rad21 band (see below). These dotted bands were found to be less significant in $cut2^{EA2}$ (Fig. 4a, right panel), suggesting that they might represent radiation-induced Rad21 cleaved products. The hyper-shifted upper band of Rad21 was abolished in a $rad3$ (ref. 13) deletion mutant, but not in $\Delta chk1$ (Fig. 4b). In $\Delta cds1$, the degree of increase in the upper band intensity was small. A Rad21 modification step might be involved in the recognition of damaged lesions that must be repaired.

Rad21 has two separase cleavage sites, Arg 179 and Arg 231 (Fig. 4c), one of which was sufficient for cell viability²². We constructed two chromosomally integrated $rad21$ mutant strains, $R179E$ and $R231E$, which contained only a single separase cleavage site. In the absence of irradiation, the cleaved Rad21 band, enhanced in $rad21^{D232S}$ due to the retarded N-end rule degradation (right panel), was altered to a higher molecular weight band (indicated by arrowhead) in the $rad21^{R231E}$ mutant. The same cleaved bands were seen in $rad21^{WT}$ and $rad21^{R179E}$. These results were consistent with the previous study²² that Arg 231 was the principal separase cleavage site.

After ultraviolet irradiation (Fig. 4d), $rad21^{WT}$ and $rad21^{R179E}$ mutant cells contained the two cleaved products (indicated by dots; see Fig. 4a) at the positions slightly different from that for the cleaved band in the absence of irradiation. However, they were not found in $rad21^{R231E}$. These results supported a hypothesis that these (dotted) bands were the radiation-induced Rad21-cleaved bands. Hyper-modified Rad21 produced upon irradiation might be the target of separase.

In short, we show that securin–separase complex has an interphase role for genome maintenance. A new repair mechanism involving securin and separase may exist. We provide evidence that the DNA damage repair role of securin–separase in interphase acts through the cleavage of cohesin Rad21 subunit, which may have to occur after damage for proper repair. Results of double mutant analysis, X-ray and γ -ray sensitivities suggested that the repair process might include DNA double-strand break repair, consistent with the original identification of a $rad21$ mutant that showed defects in double-strand break repair¹². Such a repair process is plausible when the region of DNA is large enough and requires the release of the two adhered chromatids to be repaired. In such a case, the release of cohesin in the vicinity of the lesion may facilitate recombination-based repair. The budding yeast securin/*pds1* mutant is also radiation sensitive, but the sensitivity in this case is because of the lack of the DNA damage checkpoint^{23,24}. However, securin–separase mediated repair might be hidden. Human securin

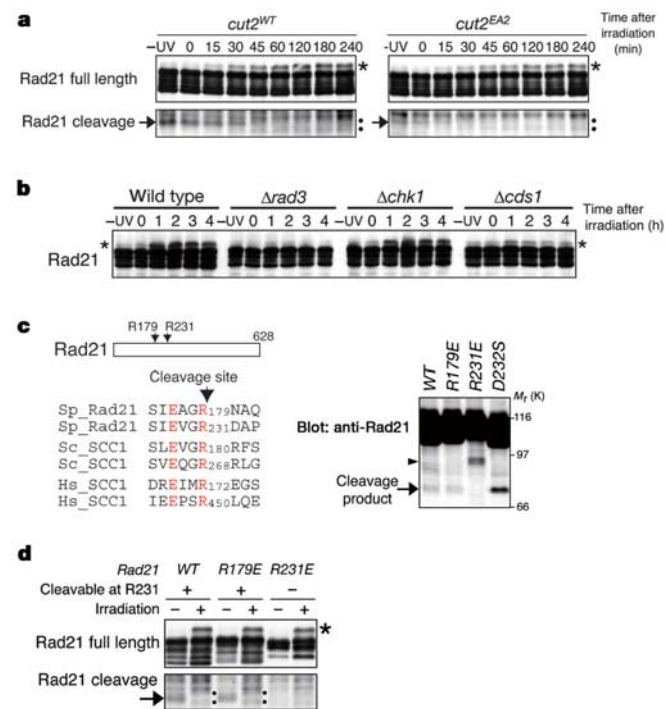


Figure 4 Rad21 cohesin is hyper-modified in response to ultraviolet irradiation.

a, Immunoblot of cell extracts prepared from asynchronous $cut2^{WT}$ and $cut2^{EA2}$ after ultraviolet (UV) irradiation (100 J m⁻²) using antibodies against Rad21. A hyper-shifted upper band (asterisk), cleaved product before irradiation (arrow) and newly cleaved products after irradiation (dots) are indicated. **b**, Formation of the hyper-shifted upper band was abolished in $\Delta rad3$. **c**, The separase cleavage sites in *S. pombe* (Sp) Rad21 are shown in comparison with those in *S. cerevisiae* (Sc) and *Homo sapiens* (Hs) (left). The cleavage products at the sites of R231 (arrow) and of R179 (arrowhead) were detected in asynchronous cultures at 26°C by anti-Rad21 antibodies (right). **d**, Immunoblot of cell extracts prepared from the cultures of $rad21^{WT}$, $rad21^{R179E}$ and $rad21^{R231E}$ cells before and 3 h after ultraviolet irradiation (100 J m⁻²). The symbols used labelled the same bands as in **a**.

interacts with p53 and modulates its transcriptional activity, but the involvement of separase is thought to be unlikely²⁵. Judging from the well-conserved cleavage of cohesin, requirement of separase in DNA repair may be general in eukaryotes.

Why is *cut2*^{EA2} normal in chromosome segregation, whereas it is hypersensitive to ultraviolet irradiation, at 26 °C? Separase was previously shown to accumulate in the mitotic nucleus, but it is sparse in the interphase nucleus⁷. Such mitosis-specific accumulation may insure that excess separase activity is available to support the fast proteolysis in anaphase. In *cut2*^{EA2}, the accumulation mechanism should be intact, but separase in the interphase nucleus after irradiation should be meagre, leading to slow repair. A similar situation may explain the condensin *cnd2-1* mutant²⁶. □

Methods

Strains, plasmids and media

All *S. pombe* strains used were derived from haploid *h*[−] 972 and *h*⁺ 975. Complete YPD and minimal EMM2 media were used. Thiamine (2 μM) was added to repress the *nmr1* promoter. The mutated *cut2* gene (designated EA2) was constructed by site-directed mutagenesis. For construction of the *cut2*^{EA2} strain, the mutated *cut2*^{EA2} gene was introduced together with the *nmr1*⁺ 3' non-coding region and *ura4*⁺ gene into the endogenous *cut2* locus of diploid strain (*h*⁺/*h*[−] *leu1/leu1* *his7/his7* *ura4/ura4* *ade6-M210/ade6-M216*). Correct integration was confirmed by genomic Southern hybridization. Resulting heterozygous diploids were sporulated, and a *Ura*⁺ haploid strain was backcrossed. The mutated *rad21* genes were cloned into the pFA6a-KanMX6 derivative plasmid, followed by the integration onto the endogenous *rad21* locus. An inducible promoter REP41 was used for moderate overexpression of *Rad21*^{WT} and *Rad21*^{REE}. The selection markers for multicopy plasmids were the *Saccharomyces cerevisiae* *LEU2* gene or the *S. pombe* *aur1*^R gene (TaKaRa). Other strains used were described previously^{9–11,16,18–20,22,27}.

Immunocytochemistry

Cell extracts were prepared using glass beads in KB buffer (25 mM Tris-HCl at pH 7.5, 15 mM EDTA, 60 mM β-glycerophosphate, 0.1% NP-40, 10% glycerol, 1 mM phenylmethylsulphonyl fluoride, 2 mg ml^{−1} pepstatin A and 1 μg ml^{−1} aprotinin) or 10% trichloroacetic acid. Cell extracts were boiled in LDS sample buffer (Invitrogen). Immunoblotting was performed using the following antibodies: anti-Cut1 (ref. 28), anti-HA 16B12 (Babco) and anti-Rad21 (T. Tomonaga, unpublished). The immunoprecipitation procedure was performed as described previously²².

Other techniques

Ultraviolet irradiation and detection of thymine dimers were performed as previously described²⁹. For quantification of dot blots, the intensity was measured by GS-250 Molecular Imager (BioRad), with calibration using serial dilutions of the 0-time-point DNA. The X-ray and γ-ray apparatuses used were a RadioFlex RF-350 (Rigaku Denki) with the dose rate of 32.6 Gy min^{−1} and a Gammacell 40 Exactor (MDS Nordion) with the dose rate of 1.245 Gy min^{−1}, respectively. For the synchronous culture, a Beckman elutriator rotor (J-6M/E) was used to centrifuge exponentially growing cells (1.5 × 10¹⁰) at 26 °C, and small cells (7 × 10⁶) were collected and grown in YPD medium at 26 °C.

Received 12 April; accepted 1 July 2004; doi:10.1038/nature02803.

- Nasmyth, K. Segregating sister genomes: the molecular biology of chromosome separation. *Science* **297**, 559–565 (2002).
- Funabiki, H. *et al.* Cut2 proteolysis required for sister-chromatid separation in fission yeast. *Nature* **381**, 438–441 (1996).
- Cohen-Fix, O., Peters, J. M., Kirschner, M. W. & Koshland, D. Anaphase initiation in *Saccharomyces cerevisiae* is controlled by the APC-dependent degradation of the anaphase inhibitor Pds1p. *Genes Dev.* **10**, 3081–3093 (1996).
- Uhlmann, F., Lottspeich, F. & Nasmyth, K. Sister-chromatid separation at anaphase onset is promoted by cleavage of the cohesin subunit Scc1. *Nature* **400**, 37–42 (1999).
- Uhlmann, F., Wernic, D., Poupart, M. A., Koonin, E. V. & Nasmyth, K. Cleavage of cohesin by the CD clan protease separin triggers anaphase in yeast. *Cell* **103**, 375–386 (2000).
- Hauf, S., Waizenegger, I. C. & Peters, J. M. Cohesin cleavage by separase required for anaphase and cytokinesis in human cells. *Science* **293**, 1320–1323 (2001).
- Kumada, K. *et al.* Cut1 is loaded onto the spindle by binding to Cut2 and promotes anaphase spindle movement upon Cut2 proteolysis. *Curr. Biol.* **8**, 633–641 (1998).
- Jager, H., Herzig, A., Lehner, C. F. & Heidmann, S. *Drosophila* separase is required for sister chromatid separation and binds to PIM and THR. *Genes Dev.* **15**, 2572–2584 (2001).
- Walworth, N., Davey, S. & Beach, D. Fission yeast chk1 protein kinase links the rad checkpoint pathway to cdc2. *Nature* **363**, 368–371 (1993).
- Carr, A. M. *et al.* Evolutionary conservation of excision repair in *Schizosaccharomyces pombe*: evidence for a family of sequences related to the *Saccharomyces cerevisiae* RAD2 gene. *Nucleic Acids Res.* **21**, 1345–1349 (1993).
- Muris, D. F. *et al.* Cloning the RAD51 homologue of *Schizosaccharomyces pombe*. *Nucleic Acids Res.* **21**, 4586–4591 (1993).
- Birkenbihl, R. P. & Subramani, S. Cloning and characterization of *rad21* an essential gene of *Schizosaccharomyces pombe* involved in DNA double-strand-break repair. *Nucleic Acids Res.* **20**, 6605–6611 (1992).

- Bentley, N. J. *et al.* The *Schizosaccharomyces pombe rad3* checkpoint gene. *EMBO J.* **15**, 6641–6651 (1996).
- Rao, H., Uhlmann, F., Nasmyth, K. & Varshavsky, A. Degradation of a cohesin subunit by the N-end rule pathway is essential for chromosome stability. *Nature* **410**, 955–959 (2001).
- Jallepalli, P. V. *et al.* Securin is required for chromosomal stability in human cells. *Cell* **105**, 445–457 (2001).
- Saka, Y., Esashi, F., Matsusaka, T., Mochida, S. & Yanagida, M. Damage and replication checkpoint control in fission yeast is ensured by interactions of Crb2, a protein with BRCT motif, with Cut5 and Chk1. *Genes Dev.* **11**, 3387–3400 (1997).
- Lindsay, H. D. *et al.* S-phase-specific activation of Cds1 kinase defines a subpathway of the checkpoint response in *Schizosaccharomyces pombe*. *Genes Dev.* **12**, 382–395 (1998).
- Stewart, E., Chapman, C. R., Al-Khodairy, F., Carr, A. M. & Enoch, T. *rqh1*⁺, a fission yeast gene related to the Bloom's and Werner's syndrome genes, is required for reversible S phase arrest. *EMBO J.* **16**, 2682–2692 (1997).
- Murakami, H. & Okayama, H. A kinase from fission yeast responsible for blocking mitosis in S phase. *Nature* **374**, 817–819 (1995).
- Yonemasu, R. *et al.* Characterization of the alternative excision repair pathway of UV-damaged DNA in *Schizosaccharomyces pombe*. *Nucleic Acids Res.* **25**, 1553–1558 (1997).
- Nakamura, T., Nagao, K., Nakaseko, Y. & Yanagida, M. Cut1/separase C-terminus affects spindle pole body positioning in interphase of fission yeast: pointed nuclear formation. *Genes Cells* **7**, 1113–1124 (2002).
- Tomonaga, T. *et al.* Characterization of fission yeast cohesin: essential anaphase proteolysis of Rad21 phosphorylated in the S phase. *Genes Dev.* **14**, 2757–2770 (2000).
- Yamamoto, A., Guacci, V. & Koshland, D. Pds1p, an inhibitor of anaphase in budding yeast, plays a critical role in the APC and checkpoint pathway(s). *J. Cell Biol.* **133**, 99–110 (1996).
- Sanchez, Y. *et al.* Control of the DNA damage checkpoint by chk1 and rad53 protein kinases through distinct mechanisms. *Science* **286**, 1166–1171 (1999).
- Bernal, J. A. *et al.* Human securin interacts with p53 and modulates p53-mediated transcriptional activity and apoptosis. *Nature Genet.* **32**, 306–311 (2002).
- Aono, N., Sutani, T., Tomonaga, T., Mochida, S. & Yanagida, M. Cnd2 has dual roles in mitotic condensation and interphase. *Nature* **417**, 197–202 (2002).
- Hirano, T., Funahashi, S., Uemura, T. & Yanagida, M. Isolation and characterization of *Schizosaccharomyces pombe cut* mutants that block nuclear division but not cytokinesis. *EMBO J.* **5**, 2973–2979 (1986).
- Uzawa, S., Samejima, I., Hirano, T., Tanaka, K. & Yanagida, M. The fission yeast *cut1*⁺ gene regulates spindle pole body duplication and has homology to the budding yeast *ESP1* gene. *Cell* **62**, 913–925 (1990).
- Esashi, F. & Yanagida, M. Cdc2 phosphorylation of Crb2 is required for reestablishing cell cycle progression after the damage checkpoint. *Mol. Cell* **4**, 167–174 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We would like to thank C. Shimoda and T. Nakamura for the *S. pombe* genomic library, T. Carr, H. Okayama and N. Walworth for strains, and T. Matsumoto for introducing X-ray and γ-ray irradiation. This study was supported by the COE Grant of the Ministry of Education, Culture, Sports, Science and Technology of Japan. K.N. was a recipient of a predoctoral fellowship from the Japan Society for Promotion of Science.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.Y. (yanagida@kozo.lif.kyoto-u.ac.jp).

Structural basis for glycosphingolipid transfer specificity

Lucy Malinina¹*, Margarita L. Malakhova²*, Alexei Teplov¹, Rhoderick E. Brown² & Dinshaw J. Patel¹

¹Structural Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA

²Hormel Institute, University of Minnesota, Austin, Minnesota 55912, USA

*These authors contributed equally to this work

Lipid transfer proteins are important in membrane vesicle biogenesis and trafficking, signal transduction and immunological presentation processes^{1–3}. The conserved and ubiquitous mammalian glycolipid transfer proteins (GLTPs) serve as potential regulators of cell processes mediated by glycosphingolipids, ranging from differentiation and proliferation to invasive

adhesion, neurodegeneration and apoptosis^{4,5}. Here we report crystal structures of apo-GLTP (1.65 Å resolution) and lactosylceramide-bound (1.95 Å) GLTP, in which the bound glycosphingolipid is sandwiched, after adaptive recognition, within a previously unknown two-layer all- α -helical topology. Glycosphingolipid binding specificity is achieved through recognition and anchoring of the sugar-amide headgroup to the GLTP

recognition centre by hydrogen bond networks and hydrophobic contacts, and encapsulation of both lipid chains, in a precisely oriented manner within a 'moulded-to-fit' hydrophobic tunnel. A cleft-like conformational gating mechanism, involving two inter-helical loops and one α -helix of GLTP, could enable the glycolipid chains to enter and leave the tunnel in the membrane-associated state. Mutation and functional analyses of residues in the glyco-

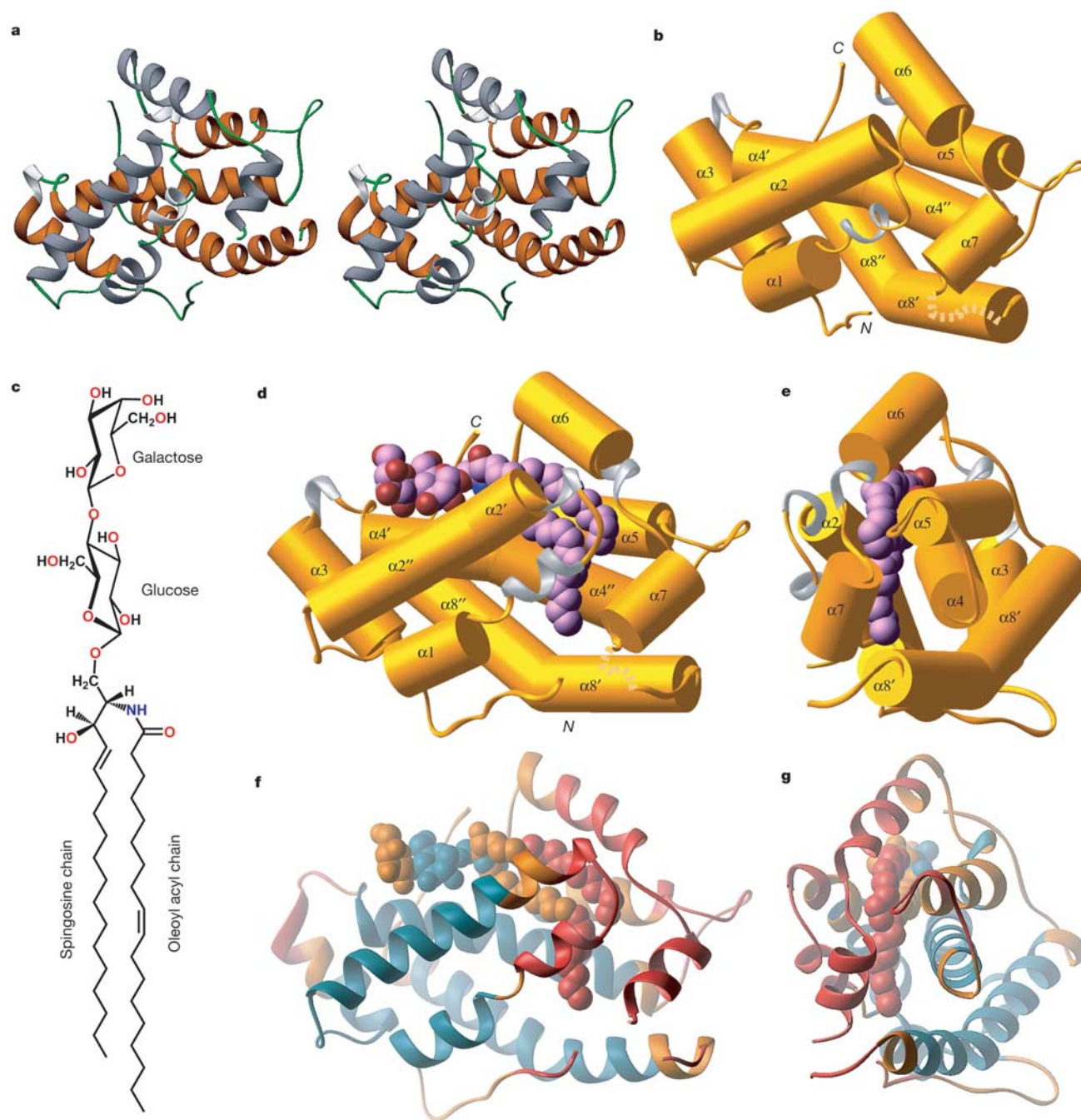


Figure 1 Overall structures of apo-GLTP and the lactosylceramide-GLTP complex. **a**, Stereo view of the apo-protein in a ribbon representation. Four α -helices on the front face are coloured grey, and four other α -helices on the back face are coloured brown. **b**, A view of apo-GLTP: α -helices (coloured gold) are shown in a cylinder representation; 3_{10} -helices (coloured silver) and loop segments (coloured gold) are shown in a ribbon representation. **c**, Chemical formula of lactosylceramide. **d, e**, Two perpendicular views of the lactosylceramide-GLTP complex, with α -helices (coloured gold) in a cylinder

representation, 3_{10} -helices (coloured silver) and loop segments (coloured gold) in a ribbon representation, and bound lactosylceramide in a space-filling representation. The bound glycolipid atoms are coloured lavender, red and blue for carbon, oxygen and nitrogen atoms, respectively. **f, g**, Views (corresponding to **d** and **e**) of the crystallographic B -factors for the lactosylceramide-GLTP complex. The B -factor scale is colour coded as follows: blue-green, $B < 34$; orange, $34 < B < 46$; red, $B > 46$.

lipid recognition centre and within the hydrophobic tunnel support a framework for understanding how GLTPs acquire and release glycosphingolipids during lipid intermembrane transfer and presentation processes.

GLTPs are soluble proteins discovered initially in the membrane-free cytosolic extract of bovine spleen, and later in a wide variety of tissues, that selectively accelerate the transfer of glycolipids between membranes^{6–8}. The 23–24-kDa GLTPs display absolute specificity for glycolipids⁹ and are highly conserved among mammals¹⁰, with some orthologues implicated in apoptosis^{11,12}. Although specific folding motifs and lipid-binding domains have been characterized in other lipid transfer/binding proteins^{13–19}, this is not true of GLTP. We have solved the structure of apo-GLTP crystal from human skin fibroblasts at 1.65 Å resolution and the corresponding lactosylceramide–GLTP complex co-crystal at 1.95 Å resolution (see Methods and Supplementary Table S1). The novel protein folding of apo-GLTP (Fig. 1a, b) is organized as a two-layer arrange-

ment of α -helices, with helices 1, 2, 6 and 7 forming the front layer, and helices 3, 8, 4 and 5 belonging to the back layer (Supplementary Fig. S1), with two long α -helices, 8 and 4, coiling around each other to form a superhelix. Such a sandwich topology generates a lactosylceramide-binding site between the α -helical layers in the lactosylceramide–GLTP complex (Fig. 1d, e; Supplementary Fig. S2a, b).

Lactosylceramide (Fig. 1c) binding to GLTP is achieved through recognition and anchoring of the hydrophilic headgroup to the protein surface, and through accommodation of the hydrophobic lipid chains within the interior of the protein. The lactosyl headgroup is anchored within the GLTP recognition centre, located on the protein surface, through a network of hydrogen-bonding (with Asp 48, Asn 52 and Lys 55 of α 2, and Tyr 207 towards the carboxy terminus) and hydrophobic (Trp 96 of α 4 stacks over the glucose ring, whereas Leu 92 contacts the galactose ring) interactions (Fig. 2a; Supplementary Fig. S3), analogous to carbohydrate–lectin

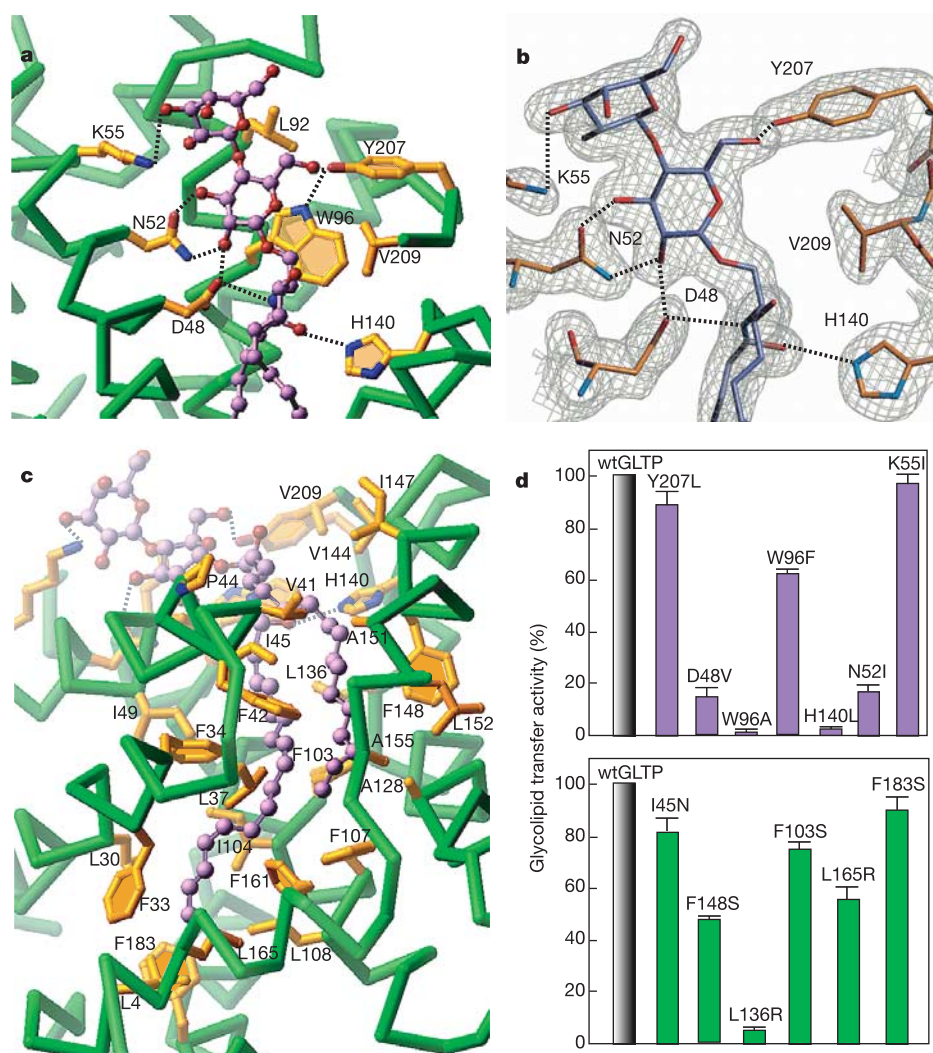


Figure 2 Intermolecular interactions in the lactosylceramide–GLTP complex. **a**, The headgroup recognition centre residues interacting with the two sugars and the ceramide amide group of bound lactosylceramide. Hydrogen bonds are shown by dashed lines. The bound glycolipid atoms are coloured lavender, red and blue for carbon, oxygen and nitrogen atoms, respectively. The GLTP C α backbone is coloured green, the side chains are shown in gold, and oxygen and nitrogen atoms are red and blue, respectively. **b**, Final $2F_o - F_c$ electron density map for the headgroup recognition centre with bound

lactosylceramide. The hydrogen bonds are indicated by dashed lines. **c**, The hydrophobic tunnel residues surrounding the longer *N*-acyl and shorter sphingoid-base chains of bound lactosylceramide. **d**, Glycolipid transfer activities for GLTPs with point mutations within the headgroup recognition centre (top) and within the hydrophobic tunnel (bottom). Error bars show standard deviations. Additional details of radiolabelled glycolipid assay conditions are provided in Methods and in Supplementary Fig. S4.

interactions^{20,21}. The ceramide amide group is oriented through a pair of hydrogen bonds (with negatively charged Asp 48 and positively charged His 140), whose alignment is facilitated by hydrophobic contacts (with Val 209) of the initial three-carbon ceramide segment (Fig. 2a; Supplementary Fig. S3). The importance of individual recognition events within the headgroup recognition centre was probed by point mutation studies, as monitored by glycolipid intervesicular transfer assays with radiolabelled

glycolipid (Supplementary Fig. S4). The experimental data support the key role of Trp 96, His 140, Asp 48 and Asn 52 for recognition of the glycosyl and ceramide amide moieties in the complex. Mutations W96A and H140L resulted in almost complete inactivation (residual activity 1–3%), whereas mutations D48V and N52I resulted in significant inactivation (residual activity about 15%) (Fig. 2d, top). The extent of inactivation is different for the W96F mutation (residual activity 63%) relative to its W96A counterpart (residual activity 1%), highlighting the importance of aromatic ring stacking over the glucose ring to recognition (Fig. 2d, top). By contrast, Tyr 207 and Lys 55, which form single intermolecular hydrogen bonds, seem to be much less crucial for recognition, because the mutations Y207L and K55I retain practically full (90–97%) activity (Fig. 2d, top). We have verified the integrity of our GLTP headgroup mutants for unanticipated structural alterations. The crystal structure of the D48V mutant shows only local rearrangements (Supplementary Fig. S5), whereas circular dichroism (CD) spectra of other mutants indicate minimal perturbation of α -helical global folds (far-ultraviolet CD; Supplementary Fig. S6, top) and of aromatic amino acid environments (near-ultraviolet CD; Supplementary Fig. S7, top).

When GLTP acquires lactosylceramide, both hydrocarbon chains of the ceramide moiety become buried within a single completely hydrophobic tunnel, lined by the side chains of nonpolar phenylalanine, leucine, isoleucine and alanine residues, together with a few valine and proline residues (Fig. 2c). Residual activities ranging from 50% to 90% were observed when individual phenylalanines were replaced by serines at positions 103, 148 and 183 (Fig. 2d, bottom), indicative of a possible additive role in creating the favourable hydrophobic environment for the nonpolar chains. Significant retention of activity (82%) was also found for the I45N mutant (Fig. 2d, bottom), located near the entrance to the channel. Replacement of leucines by arginines affected activity differently depending on their location within the channel. Thus, the L165R, mutant located towards the end of the channel, retained 54% of activity, whereas L136R, located in the interior of the channel, retained only 5% of activity (Fig. 2d, bottom). We have verified the integrity of the GLTP channel mutants by recording the far-ultraviolet and near-ultraviolet CD spectra (Supplementary Figs S6 and S7, bottom panels).

Neither the hydrophobic tunnel nor lipid chains (a $\Delta 9,10$ *cis*-double bond occurs halfway down the *N*-acyl chain) are straight in the complex (Figs 1d, e and 2c). The *N*-acyl chain extends deeper into the hydrophobic tunnel because it is longer than the adjacent sphingoid-base chain (Fig. 1c, d). The terminal segments of both chains become nearly orthogonal with respect to the long axis of the sphingoid base (Figs 1d and 2c). Our measured solvent-accessible volume of about 315 Å³ for the lipid-binding tunnel in the complex is much smaller than reported values for other lipid-binding pockets¹⁹, indicative of a very tight fit between the inserted portion of the lipid chains, with an estimated volume of about 330 Å³, and the walls of the tunnel. The measured solvent-accessible volume is further reduced to 170 Å³ in apo-GLTP, so that this residual space can no longer accommodate lipid chains.

Our structure of the lactosylceramide–GLTP complex (1.95 Å) can be compared with the structures of the ganglioside GM2–CD1b (2.8 Å)¹⁶ and sulphated-galactosylceramide–CD1a (2.15 Å)¹⁷ complexes. The recognition specificity associated with the network of hydrogen-bonding interactions involving the glycosyl and ceramide moieties distinguishes the GLTP complex (Fig. 2a) from the corresponding CD1a complex, which involved far fewer hydrogen-bonding recognition contacts¹⁷. Further, whereas the dual-chain lipid inserts into a common hydrophobic tunnel in the GLTP complex (Fig. 2c), the sphingoid-base and *N*-acyl chains insert into separate interconnecting hydrophobic pockets in the CD1a complex¹⁷. It therefore seems that unique glycolipid–protein inter-

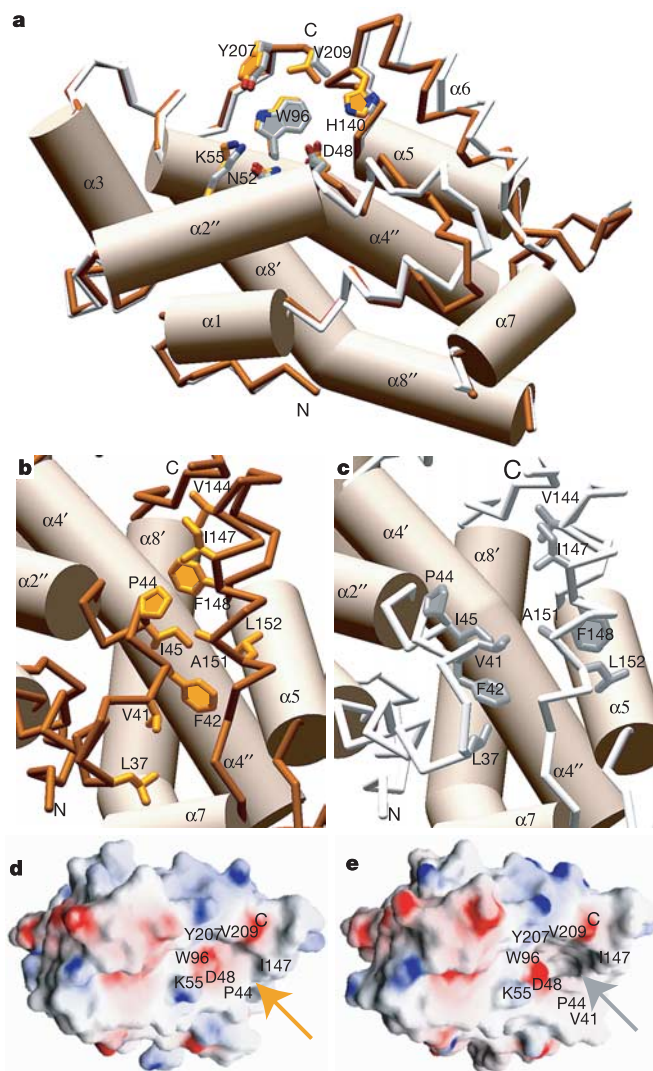


Figure 3 Comparison of apo-GLTP and lactosylceramide-bound GLTP. **a**, Superposition of apo-GLTP and lactosylceramide-bound GLTP structures, with similarly positioned α -helices shown in a cylinder representation, while 3_{10} helices and loop segments are coloured orange for free GLTP and white for lactosylceramide-bound GLTP. Note the similarity of the glycosyl recognition centre conformations (top centre) and differences in the structure of the $\alpha 1$ – $\alpha 2$ loop, and in mutual positioning of the $\alpha 6$ -helix and the N terminus of the $\alpha 2$ -helix. **b**, **c**, Detailed views of the positioning of the backbone and side chains between the $\alpha 1$ – $\alpha 2$ loop and the $\alpha 6$ -helix in apo-GLTP (**b**) and in the lactosylceramide–GLTP complex (**c**). Note the conformational rotation of Phe 148 (F148) and Leu 152 (L152) from an inwardly pointing orientation in **b** to a swung-out orientation in **c**. **d**, **e**, Electrostatic surface of apo-GLTP (**d**) and lactosylceramide-bound GLTP (**e**). The channel (indicated by a grey arrow) leading from the glycolipid recognition centre is open in the lactosylceramide–GLTP complex (**e**), but is closed (indicated by a gold arrow) in apo-GLTP (**d**). Amino acids lining the entrance to the channel are labelled.

actions are associated with distinct functional events, namely glycolipid transfer in the GLTP complex in contrast to glycolipid presentation in the CD1b¹⁶ and CD1a¹⁷ major histocompatibility complexes.

Superposition of the apo-GLTP and lactosylceramide-bound GLTP structures reveals essentially no difference in the headgroup recognition centre but provides clear evidence of localized conformational differences related to the hydrophobic channel (Fig. 3a). The differences, positioned along the front layer of GLTP, are associated with interhelical loops $\alpha 1$ – $\alpha 2$ and $\alpha 6$ – $\alpha 7$, helix $\alpha 6$ and the amino terminus of helix $\alpha 2$ (Fig. 3b, c). The conformational consequences of glycolipid acquisition are: first, bending of the $\alpha 2$ -helix; second, rearrangement of the $\alpha 1$ – $\alpha 2$ loop, accompanied by the appearance of a new 3_{10} helix near the N terminus of the $\alpha 2$ helix; third, outward displacement of the $\alpha 6$ helix by 2.7 Å; and fourth, shortening of the C terminus of the $\alpha 6$ helix that is compensated for by formation of a new 3_{10} helix (Figs 1b, d and 3a; Supplementary Figs S1, S2c and S8c). Acquisition of glycolipid by GLTP results in amino acids Pro 44, Ile 45, Val 41, Phe 42, Leu 37 and Ile 147, Phe 148, Ala 151, Leu 152, associated with separate regions, moving farther apart with respect to each other (Fig. 3b, c), while the side chains of Phe 148 and Leu 152 rotate around their C α –C β bonds to move from an inside-facing orientation in apo-GLTP (Fig. 3b) to a swung-out orientation in lactosylceramide-bound GLTP (Fig. 3c). The net result of all movements and rearrangements is an opening and expansion of a hydrophobic tunnel (Fig. 3d, e) when GLTP acquires the lactosylceramide molecule from the membrane.

It is difficult to imagine penetration of the lipid chains into the protein interior via an 'end-on digging or tunnelling mechanism' (pushing the hydrophobic walls apart to get the terminal methyl groups buried deeper and deeper). Rather, the GLTP structure most probably provides a cleft-like gate, which could open and close to let the lipid chains in and out. Indeed, the crystallographic *B*-factor distribution reveals local conformational mobility within the lactosylceramide-bound GLTP (Fig. 1f, g; highest *B*-factors in red). A comparative analysis for apo-GLTP and lactosylceramide-bound GLTP structures suggests that the loops $\alpha 1$ – $\alpha 2$ and $\alpha 5$ – $\alpha 6$, helix 6 and possibly helix 7 could form the lipid 'gate' in GLTP. The proper orientation of the glycolipid recognition centre and the 'gate' region with respect to the membrane surface might be aided by Trp 142 as well as Tyr 81, Tyr 153, Tyr 157 and Tyr 207, nearby residues known to interact favourably with membrane interfaces^{22,23}. The GLTP-membrane interface could potentially provide a favourable environment for protein conformational changes that enhance the binding and desorption of glycolipid amphiphiles from the membrane into GLTP.

To our knowledge, the results on apo-GLTP and lactosylceramide-bound GLTP are the first demonstration of a two-layered α -helical motif functioning to accommodate a glycolipid by the creation and expansion of a 'moulded-to-fit' hydrophobic tunnel (Fig. 3e) that does not pre-exist in the free protein (Fig. 3d). This architecture differs from all previously reported motifs of lipid-binding and lipid-transfer proteins, which are either α -helical with extensive disulphide crosslinking or contain extensive β -structure. Thus, GLTP represents a new emerging family of sphingolipid-binding/transfer proteins. Our results provide a potential framework for understanding how GLTPs, with their structurally conservative and conformationally flexible segments, acquire and release membrane glycolipids during lipid transport and presentation processes. □

Methods

Plasmid construction and mutagenesis

The open reading frame encoding human GLTP (NCBI GenBank accession no. AF209704) was subcloned into the pET-30 Xa/LIC expression vector (Novagen) by ligation-

independent cloning, resulting in N-terminal His tags and S-tags and a factor Xa cleavage site. Site-directed mutations were obtained with a QuickChange site-directed mutagenesis kit (Stratagene).

Protein expression and purification

Transformed BL21 (DE3) cells (Novagen) were grown in Luria–Bertani medium at 37 °C, induced with 0.1 mM isopropyl β -D-thiogalactoside and grown for a further 16–20 h at 15 °C. Purification of recombinant GLTP (rGLTP) from soluble lysate protein was accomplished by Ni-affinity chromatography. The His-tag and S-tag sequences were removed from rGLTP by using factor Xa. rGLTP was then purified by fast protein liquid chromatography size-exclusion chromatography. Selenomethionine (Se-Met)-labelled GLTP, incorporated by inhibiting the methionine biosynthesis pathway²⁴, retained full activity. Point mutants of the liganding site (Fig. 2d) had similar high yields and solubilities to those of the wild-type protein, consistent with folding to native-like conformations. In contrast, point mutations of essential residues in connecting regions between helices and far removed from the liganding site (for example W85-*cis*-P86 (Supplementary Fig. S8b), W85R), resulted in inactive protein, characterized by poor yield and very low solubility during expression and consistent with global folding changes.

Activity of rGLTP

The glycolipid transfer activity of GLTP was measured by using established radiolabelled assays. Negatively charged donor vesicles containing tritiated galactosylceramide (2 mol%) were incubated with GLTP (1 μ g) and a tenfold excess of neutral acceptor vesicles at 37 °C. Quantification of glycolipid transfer was achieved by liquid scintillation counting of acceptor vesicles recovered from DEAE minicolumns^{8,12}.

Crystallization and X-ray data collection

Crystals of GLTP were grown by the hanging-drop vapour-diffusion method. Protein samples (7–15 mg ml^{−1}) were mixed with equal amount of well solution (15–20% (w/v) PEG 3350, 50 mM KH₂PO₄ pH 4.5). For Se-Met-GLTP, 10 mM dithiothreitol was included in crystallization reagents. Single crystals appeared in 2–5 days at room temperature. GLTP-glycolipid crystal complexes were obtained by co-crystallization of protein with lactosylceramide (0.4 mM in 50% ethanol). The crystals appeared in 2–3 months. Lactosylceramide, which contained homogeneous oleoyl (C_{18:1}) acyl chains, was synthesized and purified as described previously²⁵. Crystals were transferred into well solution including 20% glycerol as cryoprotectant, then mounted in a fibre loop and flash-frozen in a cold nitrogen stream for X-ray data collection.

Structure determination and refinement

Crystallization of apo-GLTP and its selenomethionine derivative was achieved in space group *P*₂₁ with similar unit cell parameters and with one protein molecule in the asymmetric unit (Supplementary Table S1). The crystal structure of GLTP was determined by the multiwavelength anomalous dispersion method²⁶. Five of the six selenium sites were identified and refined by the SOLVE/RESOLVE package²⁷. The resulting electron density map was calculated after automatic maximum-likelihood density modification in SOLVE/RESOLVE. On the basis of this map, most of the protein chain was traced with the ARP/wARP program²⁸. The protein model was finally built through several cycles of manual model building with TURBO. The model was refined in REFMAC²⁹ at 1.65 Å resolution against the native data set to a final *R*-factor/*R*-free of 0.171/0.224 (Supplementary Table S1). The automatic procedure ARP/wARP was used to add solvent molecules to a final model. Two examples of the electron density map (Supplementary Fig. S8a, b) indicate the high quality of the final model. The distribution of secondary structure elements along the GLTP amino-acid sequence is shown in Supplementary Fig. S8c. The final model contains two disordered regions (1–7 and 168–171). Database searches with the SCOP (<http://scop.mrc-lmb.cam.ac.uk/scop>) and DALI (<http://www.ebi.ac.uk/dali>) programs revealed no structural similarity between the helical sandwich topology of apo-GLTP and deposited protein topologies.

Co-crystallization of GLTP with lactosylceramide was achieved in space group *C*₂ with one molecule in the asymmetric unit (Supplementary Table S1). The apo-GLTP structure was used as a search model in the molecular replacement method with AMoRe³⁰. A molecular replacement electron density map calculated at 1.95 Å resolution indicated the location of protein regions involved in conformational rearrangements. The amino-acid residues that belonged to these regions were removed from the initial model and retraced *de novo* with ARP/wARP. The rebuilt model was then refined with REFMAC. The difference electron density map calculated on the basis of the refined protein model clearly exhibited extra density, which was manually interpreted as a bound lactosylceramide molecule. Solvent molecules were added to the model by the automatic procedure ARP/wARP. The model was refined in REFMAC at 1.95 Å resolution to a final *R*-factor/*R*-free of 0.192/0.244. Figure 2b shows the headgroup and its interactions for a bound lactosylceramide molecule in the final 2*F*_o – *F*_c electron density map, and Supplementary Fig. S9 displays the omit 2*F*_o – *F*_c map for a ligand. The distribution of the secondary structure elements along the amino acid sequence (Supplementary Fig. S2c) shows distinctions from those observed for apo-GLTP (Supplementary Fig. S8c). The final model contains two disordered regions (1–3 and 168). The apo-GLTP and lactosylceramide-bound GLTP crystals contain a small piece of unknown hydrocarbon, acquired during the expression of protein in *Escherichia coli*, an unsurprising situation for glycosphingolipid-binding proteins¹⁹. The CASTp program (<http://cast.engr.uic.edu>) was used to estimate solvent-accessible volumes of cavities in apo-GLTP and lactosylceramide-bound GLTP.

Received 3 June; accepted 13 July 2004; doi:10.1038/nature02856.

1. Rogers, D. P. & Bankaitis, V. A. Phospholipid transfer proteins and physiological functions. *Int. Rev. Cytol.* **197**, 35–81 (2000).
2. Wirtz, K. W. A. Phospholipid transfer proteins revisited. *Biochem. J.* **324**, 353–360 (1997).
3. Zhou, D. et al. Editing of CD1d-bound lipid antigens by endosomal lipid transfer proteins. *Science* **303**, 523–527 (2004).
4. Hakomori, S. The glycosynapse. *Proc. Natl Acad. Sci. USA* **99**, 225–232 (2002).
5. Dwek, R. A., Butters, T. D., Platt, F. M. & Zitzmann, N. Targeting glycosylation as a therapeutic approach. *Nature Rev. Drug Discov.* **1**, 65–75 (2002).
6. Metz, R. J. & Radin, N. S. Purification and properties of a cerebroside transfer protein. *J. Biol. Chem.* **257**, 12901–12907 (1982).
7. Abe, A. & Sasaki, T. Purification and some properties of the glycolipid transfer protein from pig brain. *J. Biol. Chem.* **260**, 11231–11239 (1985).
8. Brown, R. E., Jarvis, K. L. & Hyland, K. J. Purification and characterization of glycolipid transfer protein from bovine brain. *Biochim. Biophys. Acta* **1044**, 77–83 (1990).
9. Yamada, K., Abe, A. & Sasaki, T. Specificity of the glycolipid transfer protein from pig brain. *J. Biol. Chem.* **260**, 4615–4621 (1985).
10. Lin, X., Mattijus, P., Pike, H. M., Windebank, A. J. & Brown, R. E. Cloning and expression of glycolipid transfer protein from bovine and porcine brain. *J. Biol. Chem.* **275**, 5104–5110 (2000).
11. Brodersen, P. et al. Knockout of *Arabidopsis* accelerated-cell-death1 encoding a sphingosine transfer protein causes activation of programmed cell death and defense. *Genes Dev.* **16**, 490–502 (2002).
12. Mattijus, P., Turcq, B., Pike, H. M., Molotkovsky, J. G. & Brown, R. E. Glycolipid intermembrane transfer is accelerated by HET-C2, a filamentous fungus gene product involved in the cell–cell incompatibility response. *Biochemistry* **42**, 535–542 (2003).
13. Tsujishita, Y. & Hurley, J. H. Structure and lipid transport mechanism of a StAR-related domain. *Nature Struct. Biol.* **7**, 408–414 (2000).
14. Roderick, S. L. et al. Structure of human phosphatidylcholine transfer protein in complex with its ligand. *Nature Struct. Biol.* **9**, 507–511 (2002).
15. Min, K. C., Kovall, R. A. & Hendrickson, W. A. Crystal structure of α -tocopherol transfer protein bound to its ligand: implications for ataxia with vitamin E deficiency. *Proc. Natl Acad. Sci. USA* **100**, 14713–14718 (2003).
16. Gadola, S. D. et al. Structure of human CD1b with bound ligands at 2.3 Å, a maze of alkyl chains. *Nature Immunol.* **3**, 721–726 (2002).
17. Zajonc, D. M., Elsiger, M. A., Teyton, L. & Wilson, I. A. Crystal structure of CD1a in complex with a sulfatide self antigen at a resolution of 2.15 Å. *Nature Immunol.* **4**, 808–815 (2003).
18. Mahfoud, R. et al. Identification of a common sphingolipid-binding domain in Alzheimer, prion, and HIV-1 proteins. *J. Biol. Chem.* **277**, 11292–11296 (2002).
19. Schubert Wright, C., Zhao, Q. & Rastinejad, F. Structural analysis of lipid complexes of GM2-activator protein. *J. Mol. Biol.* **331**, 951–964 (2003).
20. Weis, W. I. & Drickamer, K. Structural basis of lectin–carbohydrate interactions. *Annu. Rev. Biochem.* **65**, 441–473 (1996).
21. Feinberg, H., Mitchell, D. A., Drickamer, K. & Weis, W. I. Structural basis for selective recognition of oligosaccharides by DC-SIGN and DC-SIGNR. *Science* **294**, 2163–2166 (2001).
22. Wimley, W. W. & White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Struct. Biol.* **3**, 842–848 (1996).
23. Killian, J. A. & von Heijne, G. How proteins adapt to a membrane–water interface. *Trends Biochem. Sci.* **25**, 429–434 (2000).
24. Doublé, S. Preparation of selenomethionyl proteins for phase determination. *Methods Enzymol.* **276**, 523–530 (1997).
25. Li, X.-M., Momen, M. M., Brockman, H. L. & Brown, R. E. Lactosylceramide: effect of acyl chain structure on phase behavior and molecular packing. *Biophys. J.* **83**, 1535–1546 (2002).
26. Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51–58 (1991).
27. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
28. Lamzin, V. S. & Wilson, K. S. Automated refinement of protein models. *Acta Crystallogr. D* **49**, 129–149 (1993).
29. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
30. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1997).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the personnel at SBC beamline 19BM of the Advanced Photon Source beamline staff for assistance with data collection from multiwavelength anomalous dispersion; A. Serganov for technical support; X. Lin, T. Chung and H. Pike for their contributions to the cloning and expression of the recombinant human GLTP; X.-M. Li for synthesizing and purifying N-18:1 lactosylceramide; A. J. Windebank for help with DNA sequencing at the Mayo Molecular Biology Core Facility; T. Burghardt for help with recording near-ultraviolet CD spectra; and S. Vennyaminov in the Franklyn Prendergast laboratory for recording the far-ultraviolet CD spectra. This research was supported by NIH and the Hormel Foundation. Use of the ANL SBC beamlines at the APS was supported by the US Department of Energy, Office of Energy Research.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.E.B. (reb@umn.edu) or D.P. (pateld@mskcc.org). Coordinates have been deposited in the Protein Data Bank under accession codes 1SWX for apo-GLTP and 1SX6 for lactosylceramide-bound GLTP.

Structural rearrangements in the membrane penetration protein of a non-enveloped virus

Philip R. Dormitzer¹, Emma B. Nason³, B. V. Venkataram Prasad³ & Stephen C. Harrison^{1,2}

¹Department of Pediatrics, Harvard Medical School, and the Laboratory of Molecular Medicine, Children's Hospital, and

²Howard Hughes Medical Institute, 320 Longwood Avenue, Boston, Massachusetts 02115, USA

³Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA

Non-enveloped virus particles (those that lack a lipid-bilayer membrane) must breach the membrane of a target host cell to gain access to its cytoplasm. So far, the molecular mechanism of this membrane penetration step has resisted structural analysis. The spike protein VP4 is a principal component in the entry apparatus of rotavirus, a non-enveloped virus that causes gastroenteritis and kills 440,000 children each year¹. Trypsin cleavage of VP4 primes the virus for entry by triggering a rearrangement that rigidifies the VP4 spikes². We have determined the crystal structure, at 3.2 Å resolution, of the main part of VP4 that projects from the virion. The crystal structure reveals a coiled-coil stabilized trimer. Comparison of this structure with the two-fold clustered VP4 spikes in a ~12 Å resolution image reconstruction from electron cryomicroscopy of trypsin-primed virions shows that VP4 also undergoes a second rearrangement, in which the oligomer reorganizes and each subunit folds back on itself, translocating a potential membrane-interaction peptide from one end of the spike to the other. This rearrangement resembles the conformational transitions of membrane fusion proteins of enveloped viruses^{3–6}.

VP4 spikes on virions have 'head', 'body', 'stalk' and 'foot' regions (Fig. 1a), formed by the two VP4 trypsin cleavage fragments VP8* and VP5* (Fig. 1b). The VP8* fragment contains a globular domain, the 'VP8* core', which forms the head⁷. In some virus strains the VP8* core is a haemagglutination domain, which mediates initial cell attachment by binding sialic acid. Portions of both VP8* and VP5* make up the body. VP5* residues in the body are implicated in membrane penetration and integrin binding^{8–10}. The carboxy-terminal part of VP5* forms the foot, which is buried beneath the VP7 layer, anchoring the spike. Neutralizing antibodies against rotavirus bind VP8*, VP5* and VP7 and block cell entry¹¹. When present in the gut lumen, they protect against rotavirus gastroenteritis¹². Thus, knowledge of the structural basis for cell entry can inform vaccine development.

The VP5* fragment used for structure determination (VP5CT; Fig. 1b) is produced by serial chymotrypsin and trypsin cleavage of a VP4 precursor¹³. Expressed directly, the fragment is insoluble and probably misfolded. VP5CT has the authentic VP5* amino terminus at A248. VP4 cleavage to produce this N terminus is required to prime particles for cell entry and membrane interaction^{14,15}. The crystal structure of VP5CT contains rhesus rotavirus (RRV) VP4 residues N252–L523. Perfect hemihedral twinning prevented use of diffraction data beyond 3.2 Å resolution in the structure determination (see Methods and Supplementary Methods).

VP5CT is a well-ordered homotrimer that resembles a folded umbrella (Fig. 2a). The 'post' of the umbrella is a C-terminal, α -helical, triple coiled-coil. Each of the three panels comprising the 'shade' of the umbrella is an N-terminal globular domain (Fig. 2b). Each globular domain packs in a groove between the α -helices of the other two subunits on the post's negatively charged outer surface

(Fig. 2c). Just above the post, strands K, H, and G from each globular domain join in a continuously hydrogen-bonded β -annulus around the three-fold axis (Fig. 2a). At the top of the structure, propeller-like packing of tryptophan side chains (W262, strand B) creates an additional trimer contact. Between the β -annulus and the W262 propeller, a cavity (volume 390 Å³) centred on the three-fold axis leaves room for potential rearrangements. In full-length VP5*, the foot region (Fig. 1a), similar in mass to VP5CT, is attached to the coiled-coil.

The core of each globular domain is an eight-stranded anti-parallel β -sandwich (Fig. 2b, light and dark blue). Two features of potential functional importance project from its top edge: the GH and CD β -hairpins. By joining with strands K in the β -annulus, the GH hairpin clamps each globular domain in the 'folded umbrella' position. The flexible tip of the CD β -hairpin bears a sequence motif (DGE) implicated in rotavirus binding to α 2 β 1 integrins (Fig. 2b, e)¹⁰. The accessibility of the motif, which protrudes into solvent, supports the proposed receptor-binding function.

The tips of loops projecting from the bottom edge of the β -sandwich impart a hydrophobic apex to the globular domain (Fig. 2d), which may function in membrane penetration. Near the β -sandwich, portions of these loops form β -sheets F'G(H/H') and B'C'E'D' (purple and pink respectively in Fig. 2b). The unusual interposition of β -strand I (green) between these sheets creates a deep 'hydrophobic bowl'. The distal part of the F'G loop and the alphavirus 'fusion loop' have similar sequences⁸. Both loops lie at the hydrophobic tips of globular domains implicated in membrane disruption, but the domains and the loops themselves are otherwise structurally dissimilar (Supplementary Fig. 1 and Tables 2, 3)⁶. Thus, the sequence similarity probably reflects selection for hydrophobicity and flexibility in both loops, but not common ancestry. Each strand of the F'G loop has an in-register glycine-glycine pair (382–383 and 399–400, located under labels F' and G in Fig. 2b).

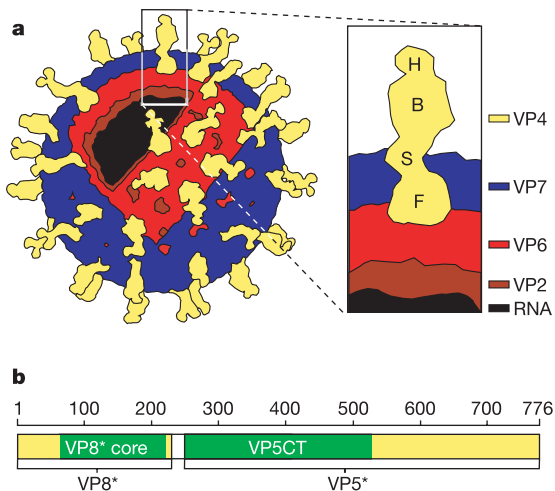


Figure 1 Rotavirus virion structure, VP4 domains and fragments. **a**, The icosahedral virion has three layers: an inner VP2 layer, which contains the genome, polymerase and capping enzyme; a middle VP6 layer; and an outer VP7 layer. To initiate infection, a 710 Å diameter subviral particle, consisting of the inner two layers and their contents, is delivered into the cytoplasm. Head (H), body (B), stalk (S) and foot (F) regions of the VP4 spikes are indicated. The approximate two-fold axis of the head and body does not correspond to any icosahedral symmetry axis. The protruding portion of the spike is within the icosahedral asymmetric unit. The drawing is based on an electron cryomicroscopy image reconstruction³⁰. **b**, Diagram illustrating the VP8* (residues M1–R231) and VP5* (A248–L776) fragments, produced by priming virions with trypsin¹⁴; the VP5CT fragment (A248–L525 to F528), produced by serially digesting soluble VP4 with chymotrypsin and trypsin¹³; and the VP8* core domain (L65–L224)⁷.

The four conserved glycines might provide a hinge, opening up the hydrophobic bowl beneath and allowing relatively extensive and flexible membrane interactions during cell entry.

How does the VP5CT structure relate to that of the VP4 spike? The region of VP5CT from Y267 to L470 neatly fits the molecular envelope of the spike body in a new, ~12 Å resolution electron cryomicroscopy image reconstruction of trypsin-primed virions (Fig. 3b, e). As Y267 and L470 hydrogen bond to each other on adjacent β -strands B and J, this region constitutes a self-contained globular domain called the VP5* antigen domain (solid colours in Fig. 3c, f). VP5CT residues outside the antigen domain must fold differently in the spike than in the trimeric crystal structure (white outlines in Fig. 3c, f).

Together, the VP5* antigen domain and the VP8* core contain all known neutralizing epitopes on VP4 (Fig. 3a, d), making them logical minimal antigens for recombinant rotavirus vaccines. The much greater conservation in VP5* suggests stronger functional constraints and correlates with heterotypic neutralization by some VP5*-specific antibodies. The VP5* antigen domain is thus a particularly promising potential immunogen. On the basis of the VP5CT structure, we have grouped VP5* neutralization escape mutations into five epitopes (Supplementary Table 1). The epitopes are solvent-exposed on both the VP5CT trimer (Fig. 2c, d) and the spike body (Fig. 3a, d). The surfaces of the trimer that lack neutralization epitopes are inaccessible on the spike, explaining their antigenic 'silence' on primed virions.

Most VP5*-specific neutralizing monoclonal antibodies map to epitope 5-1, the exposed surface of the 'shoulder' at the virion-distal end of the spike body, formed by the hydrophobic apex of the antigen domain (Fig. 3). The head covers much of the apex, but part of the F'G loop is exposed. Monoclonal antibody 2G4 selects a neutralization escape mutation in this loop at residue Q393 (ref. 8). Electron cryomicroscopy of antibody-binding fragment (FAB)-decorated virions shows that 2G4 indeed binds the shoulder¹⁶, confirming our fit of the antigen domain to the spike body. Antibody 2G4 blocks a post-attachment entry event¹¹, which is possibly domain rotation or membrane insertion (see below). Epitopes 5-1, 5-2, 5-3 and 5-4 are linked by a network of antibody competition and escape mutant cross-resistance (references in Supplementary Table 1). Epitope 5-5 is outside of this network and is mimicked by a short synthetic peptide¹⁷. It maps to the flexible CD hairpin, adjacent to the integrin-binding motif (Fig. 2b, c, e), suggesting that epitope 5-5-specific antibodies may block integrin interactions.

Image reconstructions from electron cryomicroscopy of rotavirus particles have provided good evidence for two conformational states of VP4: an uncleaved, flexible state, which does not produce visible density in averaged image reconstructions² (Fig. 4a), and a trypsin-cleaved, rigid state with approximate two-fold symmetry of the parts that project beyond the VP7 shell (Figs 3b, e and 4b)^{18,19}. Contrary to the two-fold character of the projecting part of the spikes on trypsin-primed virions, the VP5CT crystal structure reveals an unambiguous trimer (Fig. 2a, c, d). The three-fold contacts bury 3,956 Å² (25.8%) of the surface of each subunit and render the trimer resistant to dissociation by SDS¹³. Such extensive and stable contacts provide structural evidence that this region of VP4 evolved to trimerize. VP5CT thus represents a third conformational state for the projecting part of VP4 (Fig. 4c).

How can a two-fold to three-fold reorganization of VP4 occur? For the transition to take place on the virion, rather than after dissociation, each spike location must actually contain three VP4 subunits at all stages. Indeed, recently obtained image reconstructions of trypsinized rotavirus particles treated briefly at pH 11 reveal foreshortened spikes with clear three-fold character: trefoil-like features projecting slightly above the VP7 layer (B.V.V.P., manuscript in preparation). Unfolding at elevated pH appears to cause each VP4 subunit to condense so that better-ordered segments near

the foot produce strong density in the image. We conclude that two-fold clustering after trypsin cleavage leaves the third subunit still flexible, as in the uncleaved state (Fig. 4b).

Several observations support this view of VP4 organization. The spike projects from an asymmetric, angled stalk, so that its approximate two-fold axis is displaced from the centre of the foot^{18,19}. Thus, the symmetry of the foot and the projecting spike need not match. Indeed, image reconstructions of VP4 obtained by subtracting

images of spikeless particles from images of particles with spikes reveal three-fold symmetry in the foot¹⁹. The symmetry mismatch between foot and projection was previously reconciled by proposing that two VP4 subunits each contribute three similarly shaped domains to the foot. The new data suggest, instead, that the three-fold symmetry of the foot reflects the true subunit stoichiometry. Our previous biochemical data on VP5CT oligomers in solution were consistent either with trimers or with tight dimers

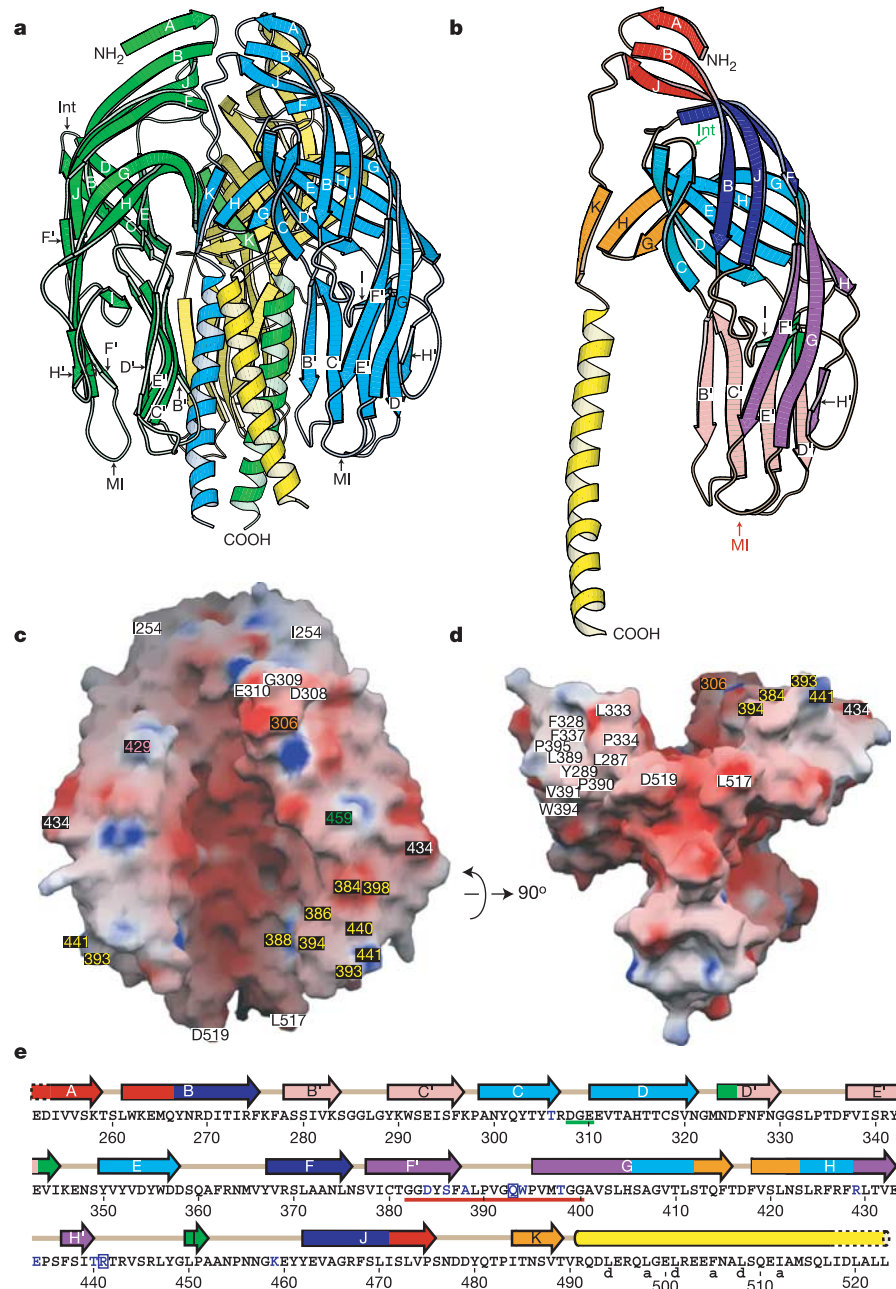


Figure 2 VP5CT structure. **a**, Ribbon diagram of VP5CT. The trimer stands 84 Å high with a 37 Å maximum radius. The green and blue subunits depict residues I254–D519; the yellow subunit depicts residues E252–L517. Int, integrin-binding motif; MI, potential membrane interaction (F'G) loop. **b**, Ribbon diagram of a single VP5CT subunit (I254–L523). The colours of the secondary structure elements match those in **e**. **c**, The VP5CT surface oriented as in **a** and coloured by electrostatic potential: blue, positive; red, negative. Surfaces formed by residues selected in neutralizing antibody escape mutants (Supplementary Table 1) are labelled with residue numbers coloured by epitope in black text boxes: 5-1, yellow; 5-2, white; 5-3, green; 5-4, pink; 5-5, orange. I254, L517 and D519 mark N- and C termini of the depicted subunits. D308, G309 and E310 form the

integrin-binding motif at the tip of the claw-like CD hairpin. **d**, The VP5CT surface, coloured as in **c** and viewed along the three-fold symmetry axis from the bottom. Surfaces formed by hydrophobic residues at the apex of the globular domain are labelled on the upper-left subunit. **e**, Secondary structure assigned to the RRV primary amino acid sequence. β-strands are arrows; the α-helix is a tube. Variable secondary structure at the N- and C termini is indicated by dashed outlines. Blue letters indicate neutralization escape mutations; escape mutations in strain RRV are boxed⁸. The green underline identifies the integrin-binding motif¹⁰. The red underline identifies the potential membrane interaction loop⁸. Residues that make hydrophobic contacts around the three-fold axis of the coiled-coil are indicated by 'a' or 'd', corresponding to heptad repeat positions.

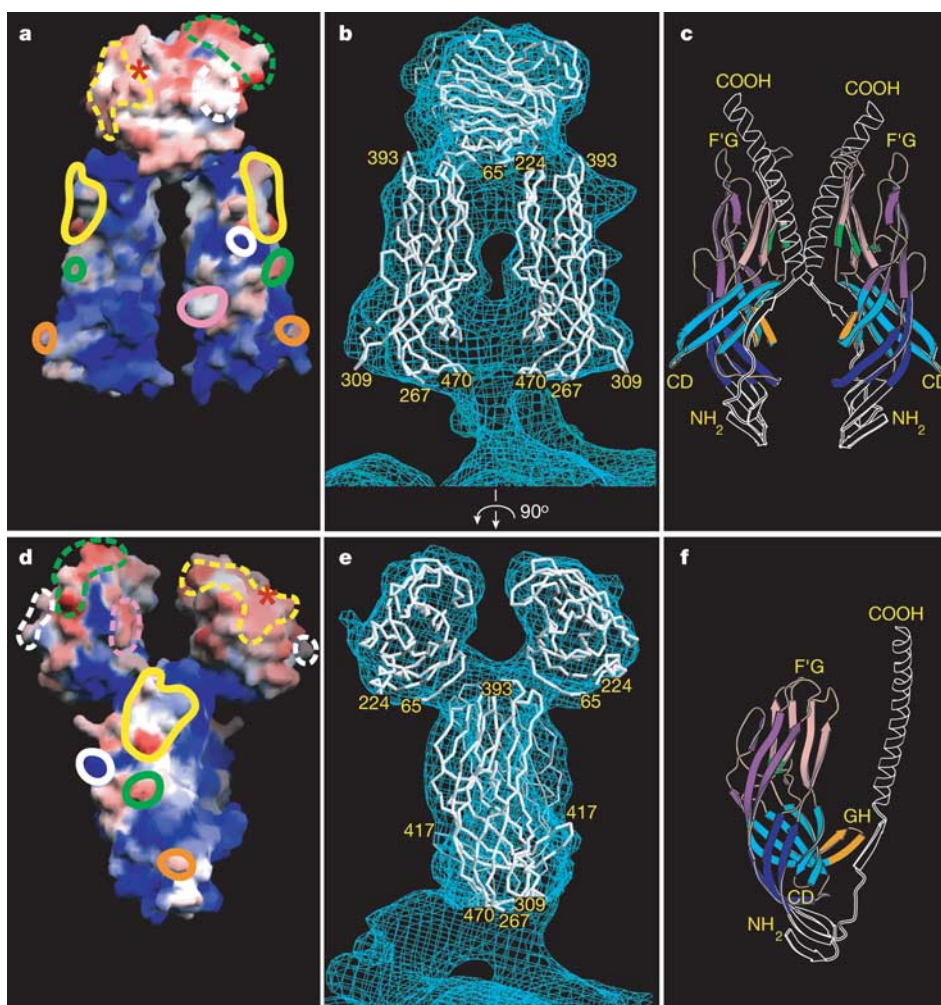


Figure 3 The VP5* antigen domain and the VP8* core fit to an electron cryomicroscopy image reconstruction. Each model in **d–f** is rotated 90° about a vertical axis relative to the corresponding model in **a–c**. **a, d**, Surface models oriented as in the spike. The surfaces are coloured by conservation in 20 P genotypes (sequences referenced in Supplementary Table 2): blue, conserved; red, variable. Neutralization epitopes are outlined: 5-1, solid yellow; 5-2, solid white; 5-3, solid green; 5-4, solid pink; 5-5, solid orange; 8-1, dashed yellow; 8-2, dashed white; 8-3, dashed green; 8-4, dashed pink. VP8* epitopes and the sialoside-binding site (red asterisk) have been described

previously⁷. **b, e**, C α traces of the VP8* core and VP5* antigen domain fit to an image reconstruction of trypsinized virions at ~12 Å resolution, contoured at 0.45 σ . Labelled residues: 65 and 224, VP8* core termini; 267 and 470, VP5* antigen domain termini; 309, glycine of the DGE integrin-binding motif (CD loop); 393, immunodominant site in the potential membrane interaction (F'G) loop; 417, tip of the GH hairpin. **c, f**, Ribbon diagrams of the VP5CT subunit, with the antigen domain oriented to fit the spike body. The antigen domain is coloured to match Fig. 2b. Residues outside the antigen domain, which must fold differently in the spike, are drawn in white outline.

that associate weakly into tetramers (dimers of dimers)¹³. At the time, the latter interpretation appeared more consistent with electron microscopy, but the crystal structure and new image reconstructions resolve the ambiguity in favour of trimers.

If just two of three VP4 subunits at each spike location cluster after trypsin cleavage, interactions with VP7 and/or VP6 must select the two that form the visible spike. The VP8* and VP5* fragments of the third, flexible molecule need not remain associated, and this subunit could even perform a different function in cell entry. Interactions with VP7 and VP6 may also arrest the progression of VP4 rearrangements at the primed stage.

In the dimeric spike, a loop (F'G) with the properties of a membrane-interacting element points away from the foot, and the GH hairpin points approximately 90° away from the dyad axis of the body (Fig. 3b, c, e, f). In the VP5CT trimer, the F'G loop points towards the C terminus of the coiled-coil, to which the stalk and foot attach, and the GH hairpin joins the β -annulus around the three-fold axis (Fig. 2a). Therefore, in the two-fold to three-fold reorganization, the VP5* antigen domain must rotate approximately 180° about an axis roughly perpendicular to the spike's

long axis (Fig. 4b, c). This rotation translocates the hydrophobic tip of the F'G loop by at least 55 Å from the shoulder of the spike towards the foot.

The two subunits of the spike body make proximal and distal dyad contacts, leaving a gap in the centre (Fig. 3b, e). The VP4 residues that participate in these contacts are outside the relatively rigid VP8* core and VP5* antigen domain and probably control their large-scale displacements. Both termini of the VP8* core are in the 'neck' between head and body, near the distal dyad contact. Both termini of the VP5* antigen domain are at the bottom of the body, adjacent to the proximal dyad contact. The antigen domain completely fills the electron microscopy envelope between the dyad contacts, leaving no unfilled density to suggest an unmodelled connecting strand. Therefore, VP8* residues outside the core (M1–V64) must form the distal dyad contact and tether head to body. During rotation of the VP5* antigen domain, disruption of the distal dyad contact would release VP8* (Fig. 4c). Similarly, VP5* residues outside the antigen domain (A248–Q266 and I471–L523 and perhaps additional C-terminal residues that connect to the stalk) must form the proximal dyad contact. In the trimeric state,

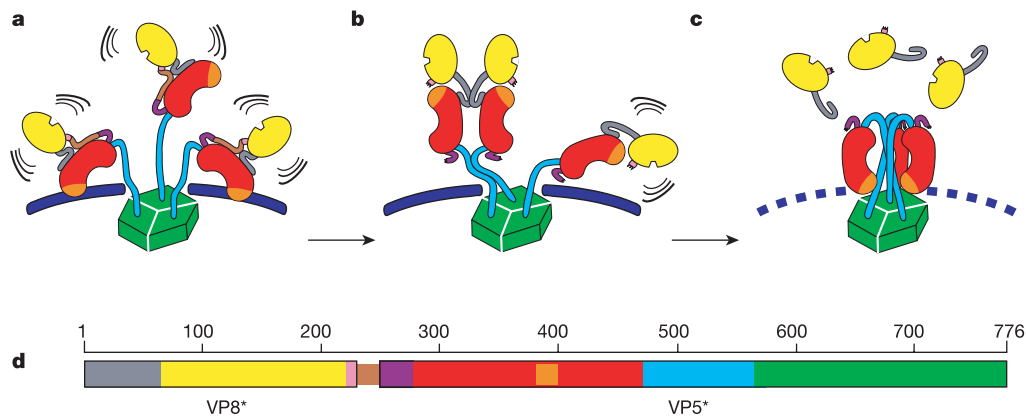


Figure 4 Model for VP4 rearrangements during priming and entry. The colours in the cartoon (a–c) match those in the linear diagram (d). The VP7 shell is blue. **a**, Uncleaved state. The wavy lines indicate flexibility. **b**, Trypsin-primed state. The activation region (brown in **a**) between VP8* and VP5* has been removed by digestion. Two clustered molecules form the visible spike. The head is the yellow oval (the VP8* core), notched at the sialoside-binding cleft. The body includes the red 'kidney' (the VP5* antigen domain), the orange cap (the potential membrane interaction loop), the purple appendage, and part

of the grey and cyan tubes. The stalk is the lower part of the cyan tubes. The foot is the green hexagonal prism. **c**, Folded-back state. This state corresponds to the VP5CT crystal structure and may form during membrane penetration. The dashed blue line indicates potential loss of VP7. **d**, Linear diagram of structural elements. The VP8* and VP5* fragments are in black boxes. Grey, residues M1–V64; yellow, L65–L224; pink, P225–R231; brown, N232–R247; purple, A248–Q266; red, Y267–T381 and A401–L470; orange, G382–G400; cyan, I471–unknown; green, unknown–L776.

these residues make up the elements of the three-fold contacts: β -sheet ABJ, β -strand K and the α -helix (Figs 2a and 3c, f). The refolding of residues in the proximal dyad contact to form the coiled-coil may drive the two-fold to three-fold reorganization.

The three distinct conformations of VP4 correspond to steps in rotavirus entry (Fig. 4a–d). When uncleaved virions are released from infected cells, three VP4 molecules with flexible stalks occupy each of 60 symmetry-equivalent positions (Fig. 4a). In the gut lumen, trypsin cleavage between VP8* and VP5* produces a first rearrangement, which primes VP5* for membrane attack (Fig. 4b). Dimeric interactions in the external portion of two primed VP4 subunits rigidify the spikes, elevating and presenting the VP8* core for cell-surface ligand binding. The third VP4 molecule remains flexible. The heads of the spike shield hydrophobic prominences on the body. We hypothesize that a subsequent, as yet unknown, entry-associated event triggers a transition to the trimeric conformation seen in the crystal structure (Fig. 4c). In this second transition, disruption of the dimer contacts releases VP8* from VP5* and unmasks the hydrophobic apex of the VP5* antigen domain, which may insert into a host-cell membrane. As the trimeric coiled-coil zips up, the antigen domain folds back against it. The resulting translocation of the potential membrane-interaction loop towards the foot could disrupt a cellular membrane. This disruption may in turn create a breach that allows a subviral particle to enter the cytoplasm, or it could lower the local calcium concentration, leading to virion uncoating and subsequent entry events. The fold-back transition could also alter interactions with receptors or directly trigger virion uncoating.

The essential properties of VP4—its capacity to adopt an inactive, precursor conformation; a primed, receptor-binding conformation; and a final, folded-back conformation—recall those of enveloped virus fusion proteins^{3–6}. Understanding these transformations will allow us to analyse neutralization mechanisms, design subunit immunogens, and elucidate molecular details of membrane perforation. □

Methods

Expression and purification of RRV VP5CT

VP5CT was produced by serial chymotrypsin and trypsin digestion of purified RRV VP4 from recombinant baculovirus-infected insect cells, as described previously¹³. Selenomethionine-substituted protein was produced by metabolic incorporation in insect cells²⁰.

Crystallization

VP5CT crystallized at 23 °C in hanging drops containing equal volumes of well solutions and a protein solution containing 9 mg ml^{−1} VP5CT, 20 mM Tris, pH 8.0, 1 mM EDTA, 0.02% sodium azide and 0.1 mM benzamidine. Selenomethionine-substituted VP5CT crystals that formed after 4 days with a well solution of 6% ethanol, 50 mM Tris pH 7.0 and 1 mM dithiothreitol were harvested after one month and flash-frozen with liquid nitrogen in a freezing buffer of 11% ethanol, 50 mM Tris, pH 7.0, 17% 2-methyl-2,4-pentenediol (MPD) and 5 mM dithiothreitol. Unsubstituted VP5CT crystals that formed after 3 weeks with a well solution of 500 mM ammonium sulphate, 5.5–6% MPD and 50 mM PIPES pH 6.5 were harvested after 1–7 months and flash-frozen in 600 mM ammonium sulphate, 22.5% MPD and 20–50 mM Tris, pH 8.0.

Structure determination

Diffraction data were scaled in space group *P4*(2)₂2, with six monomers per asymmetric unit. Starting phases were determined by multiple wavelength anomalous dispersion, based on diffraction from crystals of selenomethionine-substituted VP5CT. Thirty seleniums per asymmetric unit were located by direct methods, using SnB²¹. Landmarks provided by seleniums and bulky side chains unambiguously aligned the primary amino acid sequence to the electron density. Refinement took advantage of six-fold non-crystallographic symmetry. The final model contains no Ramachandran violations, and 76.6% of dihedral angles are in most-favoured regions. Pseudocentring and perfect hemihedral twinning in VP5CT crystals has been described previously²². Details of the structure determination are provided in Supplementary Methods, Supplementary Table 4, and Supplementary Fig. 2.

Electron cryomicroscopy

Electron micrographs of rotavirus particles (strain SA11-4F) embedded in a thin layer of vitreous ice were recorded under low electron dose conditions ($\sim 10 e^- \text{Å}^{-2}$) at $\times 50,000$ magnification on a JEOL 4000 electron microscope operated at 400 kV. Electron micrographs were digitized with a scanning interval corresponding to 2.8 Å in the object. A three-dimensional reconstruction was computed using cylindrical expansion methods^{23,24} to 12.4 Å resolution by combining 908 particles from 15 micrographs with appropriate corrections for contrast transfer function. Defocus values, determined using EMAN software²⁵, ranged from 0.8 to 3- μ m underfocus. As is conventional²³, to compute initial maps, electron microscopy data were first analysed in reciprocal space, using only 5-2-2 symmetry. To compute the final map, full icosahedral symmetry was imposed by three-fold real-space averaging after confirming the presence of three-fold symmetry in initial maps. The protruding portion of the VP4 spike was computationally extracted from the final density map. Crystal structures were fit to the electron cryomicroscopy map by a combination of automated searching using Situs²⁶ and manual fitting.

Illustration

Figures were made using O²⁷, Molscript²⁸, Grasp²⁹, Photoshop (Adobe Systems) and Illustrator (Adobe Systems).

Received 10 April; accepted 14 July 2004; doi:10.1038/nature02836.

- Parashar, U. D., Hummelman, E. G., Bresee, J. S., Miller, M. A. & Glass, R. I. Global illness and deaths caused by rotavirus disease in children. *Emerg. Infect. Dis.* **9**, 565–572 (2003).
- Crawford, S. E. *et al.* Trypsin cleavage stabilizes the rotavirus VP4 spike. *J. Virol.* **75**, 6052–6061 (2001).
- Ruigrok, R. W. *et al.* Studies on the structure of the influenza virus haemagglutinin at the pH of membrane fusion. *J. Gen. Virol.* **69**, 2785–2795 (1988).

4. Weissenhorn, W. *et al.* The ectodomain of HIV-1 env subunit gp41 forms a soluble, alpha-helical, rod-like oligomer in the absence of gp120 and the N-terminal fusion peptide. *EMBO J.* **15**, 1507–1514 (1996).
5. Modis, Y., Ogata, S., Clements, D. & Harrison, S. C. Structure of the dengue virus envelope protein after membrane fusion. *Nature* **427**, 313–319 (2004).
6. Gibbons, D. L. *et al.* Conformational change and protein-protein interactions of the fusion protein of Semliki Forest virus. *Nature* **427**, 320–325 (2004).
7. Dormitzer, P. R., Sun, Z.-Y. J., Wagner, G. & Harrison, S. C. The rhesus rotavirus VP4 sialic acid binding domain has a galectin fold with a novel carbohydrate binding site. *EMBO J.* **21**, 885–897 (2002).
8. Mackow, E. R. *et al.* The rhesus rotavirus gene encoding protein VP3: location of amino acids involved in homologous and heterologous rotavirus neutralization and identification of a putative fusion region. *Proc. Natl Acad. Sci. USA* **85**, 645–649 (1988).
9. Dowling, W., Denisova, E., LaMonica, R. & Mackow, E. R. Selective membrane permeabilization by the rotavirus VP5* protein is abrogated by mutations in an internal hydrophobic domain. *J. Virol.* **74**, 6368–6376 (2000).
10. Graham, K. L. *et al.* Integrin-using rotaviruses bind $\alpha 2\beta 1$ integrin $\alpha 2$ I domain via VP4 DGE sequence and recognize $\alpha X\beta 2$ and $\alpha V\beta 3$ by using VP7 during cell entry. *J. Virol.* **77**, 9969–9978 (2003).
11. Ruggeri, F. M. & Greenberg, H. B. Antibodies to the trypsin cleavage peptide VP8 neutralize rotavirus by inhibiting binding of virions to target cells in culture. *J. Virol.* **65**, 2211–2219 (1991).
12. Offit, P. A., Shaw, R. D. & Greenberg, H. B. Passive protection against rotavirus-induced diarrhea by monoclonal antibodies to surface proteins vp3 and vp7. *J. Virol.* **58**, 700–703 (1986).
13. Dormitzer, P. R., Greenberg, H. B. & Harrison, S. C. Proteolysis of monomeric recombinant rotavirus VP4 yields an oligomeric VP5* core. *J. Virol.* **75**, 7339–7350 (2001).
14. Arias, C. E., Romero, P., Alvarez, V. & Lopez, S. Trypsin activation pathway of rotavirus infectivity. *J. Virol.* **70**, 5832–5839 (1996).
15. Gilbert, J. M. & Greenberg, H. B. Cleavage of rhesus rotavirus VP4 after arginine 247 is essential for rotavirus-like particle-induced fusion from without. *J. Virol.* **72**, 5323–5327 (1998).
16. Tihova, M., Dryden, K. A., Bellamy, A. R., Greenberg, H. B. & Yeager, M. Localization of membrane permeabilization and receptor binding sites on the VP4 hemagglutinin of rotavirus: implications for cell entry. *J. Mol. Biol.* **314**, 985–992 (2001).
17. Taniguchi, K. *et al.* Identification of cross-reactive and serotype 2-specific neutralization epitopes on VP3 of human rotavirus. *J. Virol.* **62**, 2421–2426 (1988).
18. Shaw, A. L. *et al.* Three-dimensional visualization of the rotavirus hemagglutinin structure. *Cell* **74**, 693–701 (1993).
19. Yeager, M., Berriman, J. A., Baker, T. S. & Bellamy, A. R. Three-dimensional structure of the rotavirus haemagglutinin VP4 by cryo-electron microscopy and difference map analysis. *EMBO J.* **13**, 1011–1018 (1994).
20. Liemann, S., Chandran, K., Baker, T. S., Nibert, M. L. & Harrison, S. C. Structure of the reovirus membrane-penetration protein, $\mu 1$, in a complex with its protector protein, $\sigma 3$. *Cell* **108**, 283–295 (2002).
21. Weeks, C. M. & Miller, R. The design and implementation of SnB v2.0. *J. Appl. Crystallogr.* **32**, 120–124 (1999).
22. Padilla, J. E. & Yeates, T. O. A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning. *Acta Crystallogr. D* **59**, 1124–1130 (2003).
23. Crowther, R. A. Procedures for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs. *Phil. Trans. R. Soc. Lond. B* **261**, 221–230 (1971).
24. Lawton, J. A. & Prasad, B. V. V. Automated software package for icosahedral virus reconstruction. *J. Struct. Biol.* **116**, 209–215 (1996).
25. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).
26. Wriggers, W., Milligan, R. A. & McCammon, J. A. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* **125**, 185–195 (1999).
27. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for binding protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
28. Kraulis, J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
29. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from interfacial and thermodynamic properties of hydrocarbons. *Prot. Struct. Funct. Genet.* **11**, 281–296 (1991).
30. Yeager, M., Dryden, K. A., Olson, N. H., Greenberg, H. B. & Baker, T. S. Three-dimensional structure of rhesus rotavirus by cryoelectron microscopy and image reconstruction. *J. Cell Biol.* **110**, 2133–2144 (1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Babyonyshev for technical assistance; T. Yeates for help in analysing the crystal twinning disorder; H. Greenberg for cloned genes and recombinant baculoviruses; E. Vogan for help with data collection and analysis; and the staff of Advanced Photon Source beamline ID-19 (Argonne National Laboratory) and Cornell High Energy Synchrotron Source beamlines F1 and A1. We acknowledge the use of electron cryomicroscopy facilities at the National Center for Macromolecular Imaging funded by NIH at Baylor College of Medicine. This work was supported by an NIH grant and an Ellison Medical Foundation New Investigator in Global Infectious Diseases award to P.R.D., by an NIH grant to B.V.V.P., and by an NIH grant to S.C.H., who is a Howard Hughes Medical Institute Investigator.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.R.D. (dormitze@crystal.harvard.edu). Coordinates and observed structure factors have been deposited in the Protein Data Bank under accession code 1SLQ.

The native architecture of a photosynthetic membrane

Svetlana Bahatyrova^{1*}, Raoul N. Frese^{1,2*}, C. Alistair Siebert³, John D. Olsen³, Kees O. van der Werf¹, Rienk van Grondelle², Robert A. Niederman⁴, Per A. Bullough³, Cees Otto¹ & C. Neil Hunter³

¹Biophysical Techniques Group, Department of Science & Technology, BMTI, MESA⁺, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

²Biophysics, Faculty of Sciences, Vrije Universiteit Amsterdam, de Boelelaan 1081, 1081 HV, The Netherlands

³Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, UK

⁴Rutgers University, Department of Molecular Biology and Biochemistry, Piscataway, New Jersey 08854, USA

* These authors contributed equally to this work

In photosynthesis, the harvesting of solar energy and its subsequent conversion into a stable charge separation are dependent upon an interconnected macromolecular network of membrane-associated chlorophyll–protein complexes. Although the detailed structure of each complex has been determined^{1–4}, the size and organization of this network are unknown. Here we show the use of atomic force microscopy to directly reveal a native bacterial photosynthetic membrane. This first view of any multi-component membrane shows the relative positions and associations of the photosynthetic complexes and reveals crucial new features of the organization of the network: we found that the membrane is divided into specialized domains each with a different network organization and in which one type of complex predominates. Two types of organization were found for the peripheral light-harvesting LH2 complex. In the first, groups of 10–20 molecules of LH2 form light-capture domains that interconnect linear arrays of dimers of core reaction centre (RC)–light-harvesting 1 (RC–LH1–PufX) complexes; in the second they were found outside these arrays in larger clusters. The LH1 complex is ideally positioned to function as an energy collection hub, temporarily storing it before transfer to the RC where photochemistry occurs: the elegant economy of the photosynthetic membrane is demonstrated by the close packing of these linear arrays, which are often only separated by narrow ‘energy conduits’ of LH2 just two or three complexes wide.

Photosynthetic purple bacteria can contain two types of complex, RC–LH1 and LH2, with both light-harvesting complexes comprising roughly circularly arranged α -helices with bound carotenoid and bacteriochlorophyll (Bchl) pigments^{3–6}. To investigate the functionally crucial organization of these complexes, native photosynthetic membranes from the wild-type purple bacterium *Rhodospirillum rubrum* were imaged by atomic force microscopy (AFM), a technique that allows the topography of biological samples to be acquired in buffer solution at room temperature and under normal pressure. Figure 1a shows a cluster of several membrane patches, each of a size approximating to the surface area of an intracytoplasmic membrane vesicle (chromatophore). The bright areas represent photosynthetic complexes; even at this low magnification this remarkable view of native photosynthetic membranes shows that they are composed, at least in part, of linear arrays of dimeric complexes. This arrangement was mirrored in all the membrane patches we examined, and a gallery of additional images is shown in Supplementary Fig. S1.

Figure 1b clearly shows how the light-harvesting and photochemical functions of a membrane are apportioned, and reveals the arrangement of photosynthetic complexes. Two types can be seen: large circular complexes with a bright and therefore

protruding central protein, and smaller rings with no central density and a diameter of about 7 nm. The antenna complexes were identified by comparison with AFM data compiled on purified LH2, LH1 and RC-LH1 complexes^{7–9} and the three-dimensional structures for the LH2 and LH1-RC-PufX complexes^{1–4}. The larger features are RC-LH1-PufX complexes, each about 12 nm in diameter and comprising an LH1 ring surrounding a central bright region representing a RC complex. These RC-LH1-PufX complexes are usually dimeric, and the dimers are arranged in rows such as those indicated (red arrows) in Fig. 1b. Typically, rows of up to six dimers of RC-LH1-PufX complexes are found, forming a domain of 240–360 Bchl molecules. This is in close agreement with singlet-singlet annihilation experiments on a mutant containing only the RC-LH1-PufX complex that established domain sizes of 330 or more LH1 Bchl molecules at room temperature¹⁰.

Rather than being interspersed randomly throughout the membrane, many of the LH2 complexes are clustered in regions between the rows of RC-LH1-PufX dimers, as outlined in Fig. 1c. Typically these regions consist of 10–20 LH2 complexes representing 270–540 Bchl molecules; the same order of magnitude of connected Bchl molecules (365 or more) was deduced from singlet-singlet annihilation measurements of mutant LH2-only *R. sphaeroides* membranes¹⁰. We suggest that the LH2 complexes situated between rows of RC-LH1-PufX dimers form a relatively invariant complement of light-harvesting antenna, and that the arrangement of complexes depicted in Fig. 1c represents the basic requirement for the efficient harvesting, transmission and trapping of light energy in this bacterium. However by itself this would not account for all the LH2 present; a comparison of the starting absorbance spectrum with one estimated from the number of LH2 and LH1 rings in Figs 1 and 2 indicates that up to 50% of the expected LH2 has not been accounted for in such images. The circled region in Fig. 2a shows that there are other regions consisting largely of LH2 that are not ‘sandwiched’ between rows of RC-LH1-PufX complexes. Individual LH2 rings can be discerned within the higher-magnification image in Fig. 2b, such as the one indicated with an

arrow. These LH2-enriched domains could represent the variable antenna, known for many years to form in response to a lowered light intensity¹¹. In addition to making contact with each other, some of the LH2 rings are in close physical association with the RC-LH1-PufX complexes (see upper green arrow in Fig. 1b), facilitating the transfer of excitation energy from LH2 to LH1 and thence to the RC. Even at this intermediate level of detail, the physical and organizational basis for harvesting, transferring and using light energy is clearly evident.

A model of the bacterial photosynthetic membrane has been formulated in Fig. 1c (inset) by arranging the structural information available for the individual photosynthetic components^{2–4} in a manner consistent with the AFM data. These direct images of an intact photosynthetic membrane resemble one of the models proposed in 1976 (ref. 12) on the basis of fluorescence-quenching experiments conducted on *R. sphaeroides* membranes. From a quantitative point of view the model in Fig. 1c explains why fluorescence-quenching studies of *R. sphaeroides* membranes concluded that there were up to 3,000 light-harvesting Bchl molecules connected for energy transfer¹³. This work also estimated that a single photosynthetic membrane vesicle would contain 30 RCs, in rough agreement with the number observed in the membrane patches in Figs 1 and 2. The AFM data suggest that 100 or more LH2 molecules would form the dominant antenna within such a vesicle, partitioned between the linear rows of RC-LH1-PufX dimers (Fig. 1c) and the LH2-enriched areas in Fig. 2b. There is extensive physical continuity between individual LH2 complexes (27 Bchl molecules for each ring) and between LH2 and RC-LH1-PufX dimers (60 LH1 Bchl molecules per dimer). Thus, the LH2 rings seem to cooperate to form an extended array for collecting photons, and, once transferred ‘downhill’ to a RC-LH1-PufX dimer, the excited state can hop the relatively large distance of about 3.5 nm from LH1 Bchl molecules to the special pair of Bchl molecules within the RC¹⁴, thereby eliciting conversion to photochemical energy¹⁵. It is possible that the linear arrays of RC-LH1-PufX complexes cooperate in the overall process of energy trapping, because if any particular RC is already under-

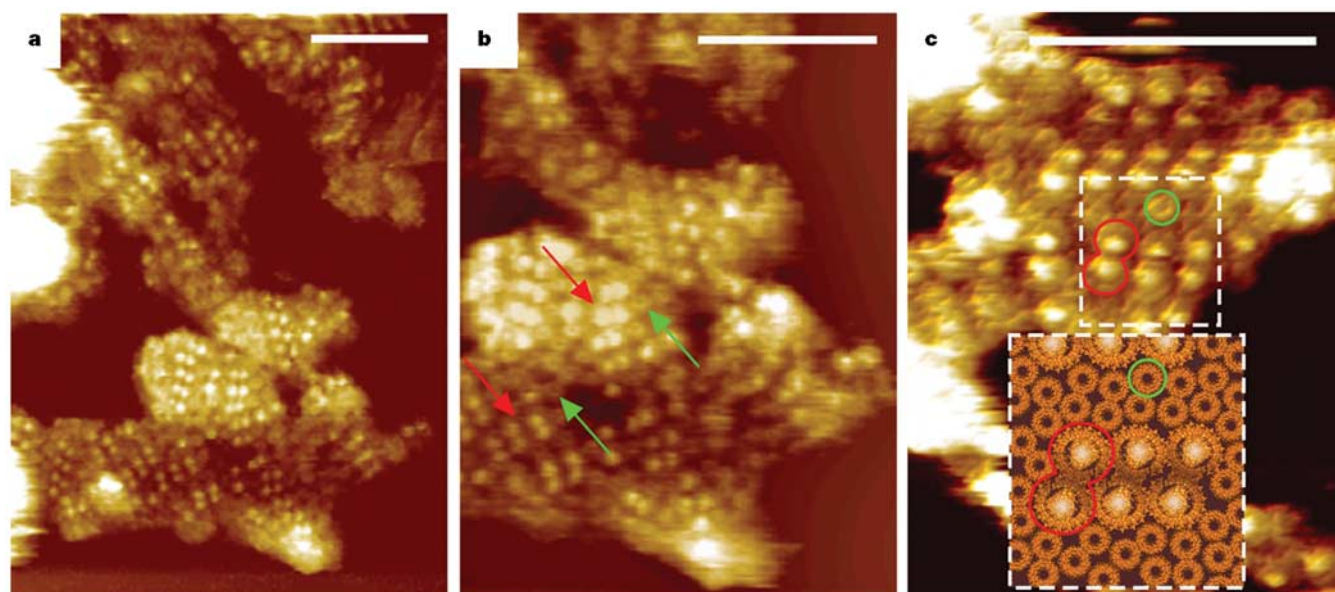


Figure 1 AFM of native photosynthetic membranes. **a**, Large-scale view of several membrane fragments. **b**, Higher-magnification view showing a region of dimeric RC-LH1-PufX core complex arrays (red arrows) and associated LH2 complexes (green arrows). **c**, Three-dimensional view of core complex arrays surrounded by LH2 complexes. The inset at the bottom is a representation of the region denoted by the

dashed box in the centre, using model structures derived from atomic resolution data^{2–4}. A typical RC-LH1-PufX dimer is delineated in both images by a red outline and a representative LH2 complex by a green circle. Scale bar, 100 nm in all panels. For all images the *z* range is 6 nm (from darkest to lightest).

taking photochemical charge separation and is thus unavailable for receiving excitation energy from its LH1 ring (in a 'closed' state), the LH1 excitation can migrate along a succession of such dimers until an 'open' RC is reached. This arrangement of RC–LH1–PufX dimers might also provide a structure with the largest number of effective connections between LH1 rings. In addition, it becomes clear why mutants lacking LH2 form tubular membranes containing linear rows of dimers aligned in parallel^{9,16}; this is a natural consequence of removing the LH2 complexes that normally separate rows of RC–LH1–PufX dimers.

The AFM was used to examine a small area of membrane containing only a few photosynthetic complexes; for clarity the data are represented in three dimensions (Fig. 3). The rows of dimeric RC–LH1–PufX complexes, which are the most prominent

features in low-magnification topographs of photosynthetic membranes (Fig. 1a, b), can now be seen in greater detail. Each dimer in Fig. 3 is about 23 nm across, corresponding to the RC–LH1–PufX dimer complexes about 20 nm in width previously revealed by negative-stain electron microscopy⁹. The central protruding feature is about 4 nm above the lipid bilayer, which corresponds to the H subunit of the RC⁸. It is therefore the cytoplasmic face of the membrane that lies uppermost and the periplasmic face that has adhered to the mica substrate. Inspection of the RC–LH1–PufX dimers shows that in some cases the bright central density is missing. This has been seen before in AFM images of bacterial RCs, and it arises when the AFM tip dislodges an extrinsic subunit, revealing the underlying L and M subunits^{8,17}. As seen in Fig. 1, some of the LH2 complexes make contact with RC–LH1–PufX complexes, at points indicated by the arrows in Fig. 3, fulfilling their role not only as gatherers of relatively high-energy excitations but also as energy conduits to the lower-energy LH1 complex surrounding the RC. Although LH2 complexes associate mainly with other LH2 complexes, some can be found singly, for example sandwiched between two RC–LH1–PufX complexes (Fig. 3, circle). At the contact points between LH2 and RC–LH1–PufX complexes, the distance between the B850 Bchl molecules of LH2 and the B875 Bchl molecules of LH1 was calculated to be between 2.7 and 3.2 nm on the basis of the energy transfer time constant, which is 3 ps (ref. 18). Remarkably, the subunit structure of both LH2 and LH1 light-harvesting rings emerges, even in the native membrane environment. The LH2 rings such as those marked by asterisks in Fig. 3 seem to be composed of nine units *in vivo*, which is consistent with previous structural data^{3,5}.

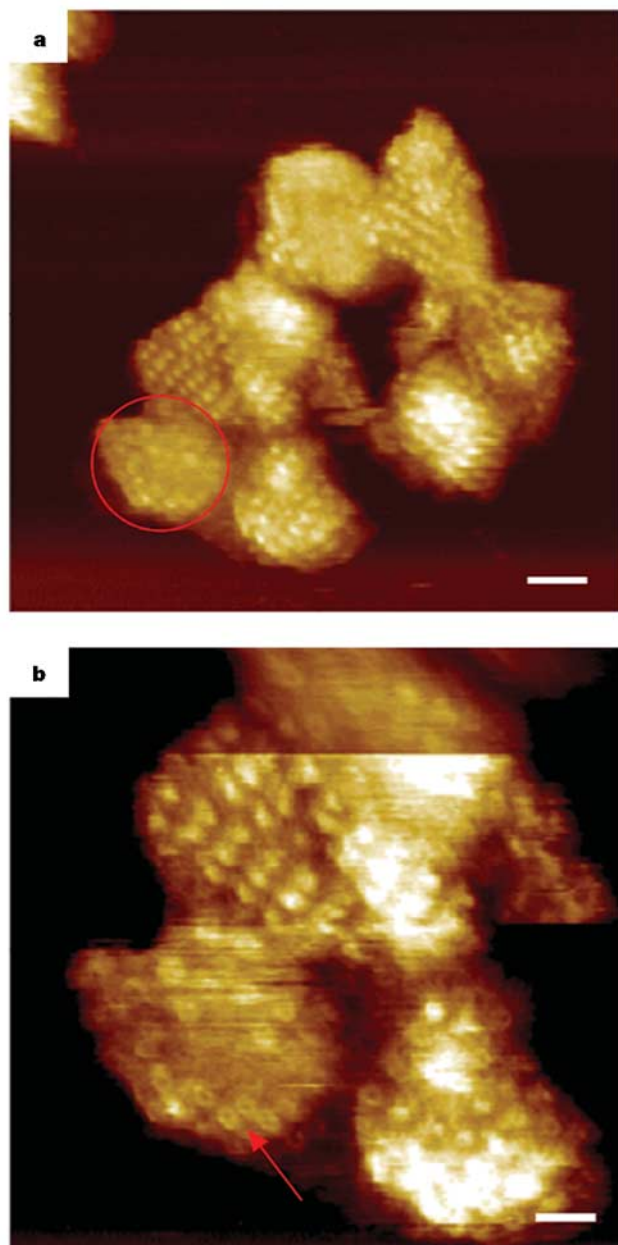


Figure 2 Membrane patches showing two types of arrangement of photosynthetic complexes. **a**, The circled region is composed mainly of LH2 complexes. **b**, Higher-magnification image of the same membrane patch in which an arrow points to an LH2 ring within the LH2-only domain. This higher-resolution scan clearly shows that there are no core complexes in these regions. Scale bar, 50 nm (**a**); 25 nm (**b**).

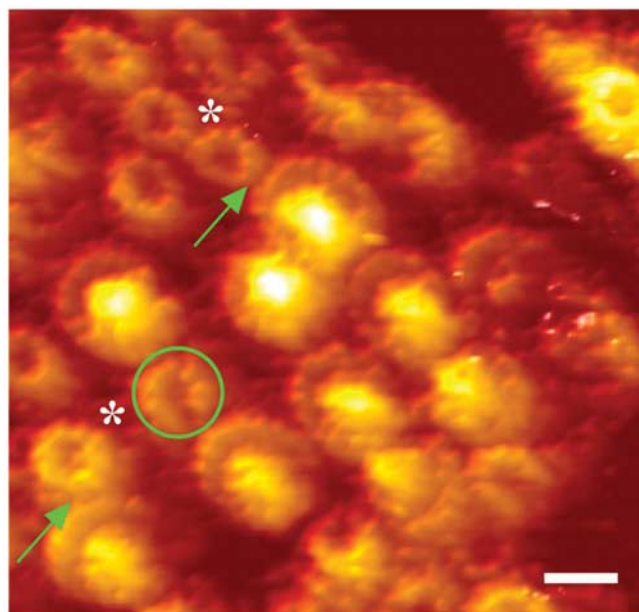


Figure 3 Three-dimensional representation of a small region of membrane showing RC–LH1–PufX core complex dimers and monomers with associated LH2 complexes. Contact points for energy transfer between LH2 and RC–LH1–PufX complexes are indicated by green arrows. LH2 rings marked by asterisks are composed of nine units. The LH2 complex in the green circle is sandwiched between two RC–LH1–PufX complexes. The average tilt of seven LH1 complexes is 4.8°; the average height of LH1 above the lipid membrane is 1.4 ± 0.3 nm (mean $\pm$ s.d.). The maximum subunit height is 1.8 ± 0.3 nm ($n = 7$) and the minimum subunit height is 1.1 ± 0.1 nm ($n = 7$). The average tilt of three LH2 complexes is 3.8°. For LH2 the average height of LH2 rings above the lipid membrane is 1.5 ± 0.2 nm ($n = 11$). For three tilted rings the maximum subunit height is 1.7 ± 0.1 nm and the minimum subunit height is 1.2 ± 0.1 nm. The average height of the reaction centre H subunit above the lipid membrane is 3.7 ± 0.3 nm ($n = 9$). Scale bar, 10 nm.

We also investigated the disposition of individual photosynthetic complexes with respect to the native membrane bilayer, information that has escaped the attention of crystallographic methods. Some, but not all, LH2 complexes seem to be slightly tilted in the native membrane, estimated to be between 3° and 4° from the vertical. This tilting is a surprise, notwithstanding the fact that LH2 is tilted in two-dimensional crystals reconstituted from detergent-purified LH2 complexes^{5,19}. This phenomenon is not confined to the LH2 complex; close inspection of seven LH1 rings within a series of native RC–LH1–PufX dimers reveals an average 4.8° tilt of each LH1 ring towards the monomer–monomer interface. We cannot determine whether the enclosed RC is also tilted, although we note that purified, detergent-solubilized RCs crystallized in a cubic lipidic phase lie 11° from the vertical²⁰. This small degree of tilt in each RC–LH1–PufX complex obscures the central part of each dimer from analysis, preventing the identification of the PufX polypeptide, which is thought to be somewhere in this region^{9,21}. It is not clear whether the tilting of LH2 and RC–LH1–PufX complexes has any functional significance, but the mutual inward tilt of the RC–LH1–PufX complexes indicates that it might be a consequence of the dimerization process. The tilting of these photosynthetic complexes might be accompanied by a slight degree of buckling of the lipid bilayer in the immediate vicinity of the complex and could arise as a consequence of applying the sample to the mica surface, which imposes a flattened profile on a piece of membrane that was originally curved. This could be exacerbated in LH2-rich regions, which are expected to be particularly curved²². An interaction with the mica surface could also account for the very low lateral mobility of the photosynthetic complexes in the membrane patches that we observed. This point is illustrated by Supplementary Fig. S2, which displays four of a sequence of ten consecutive images obtained from the same membrane patch. The reproducibility of the data is remarkable, even on the scale of a single LH2 complex. A similar lack of lateral mobility was noted previously²³ in a study of the bacterial ATP synthase in a reconstituted lipid bilayer; there was remarkably low diffusion in comparison with eukaryotic membranes. In the same study, those molecules that formed close associations with others did not tend to move and some of the isolated mobile molecules ceased moving when they associated with larger groups. We suggest that the close packing of many of the light-harvesting complexes probably prevents the lateral diffusion of individual components.

This study has directly revealed the organization of bacterial photosynthetic membranes. Further questions can now be addressed: for example, how does the increased level of LH2 in low-light-grown cells affect this architecture? Where is the other component of the cyclic electron transfer chain, the cytochrome *bc*₁ complex, located in the cell? Where is the ATP synthase? How is this simple yet effective photosynthetic membrane assembled? Further work will be necessary to answer these questions. Only small amounts of the cytochrome *bc*₁ complex were detected by western blot analysis of the membrane patches (results not shown), and there could be a potential problem in detecting cytochrome *bc*₁ complexes by AFM, because the very recently determined structure of this complex from the photosynthetic bacterium *Rhodobacter capsulatus* reveals relatively little surface topology on the cytoplasmic side of the membrane²⁴. In the case of the ATP synthase it has been reported that intracytoplasmic membrane vesicles of *R. capsulatus* contain on average only one F₀F₁-ATP synthase. Indeed, 37% of the chromatophore vesicles had no ATP synthase at all²⁵.

In view of the highly organized arrangements of photosynthetic complexes seen in recent electron microscopic investigations of bacterial and plant membranes^{9,26}, in which crystallographic order is demonstrated, our results present a slightly more chaotic landscape. A high degree of order is clearly not essential for transmitting excitation energy; the bacterial photosynthetic membrane fulfils the

basic requirements of being physically extensive to maximize the likelihood of harvesting photons, while fostering multiple contacts between its light-harvesting components so that energy can migrate between complexes. In the context of its energy transfer function this is a very robust architecture, because associations between light-harvesting rings place few demands on the contact sites as long as the distances between rings are minimized. Thus, the two-dimensional organization we have revealed here by AFM will always present multiple possibilities and pathways for the fast and efficient transfer and trapping of energy. □

Methods

Bacterial growth

Cells of *R. sphaeroides* NCIB 8253 were grown photosynthetically at moderate light intensity (500 W m⁻² for 18–20 h) and then switched to high light intensity (825 W m⁻²) for 4 h. Intracytoplasmic membrane vesicles with an LH1/LH2 molar ratio of 0.78 were isolated by rate-zonal sucrose density gradient centrifugation²⁷. Membranes were pelleted by ultracentrifugation at 100,000g for 4 h and resuspended with gentle homogenization in 50 mM HEPES buffer at pH 8 containing 0.03% β-dodecyl maltoside (buffer A) to 16 A₈₇₅ units cm⁻¹ ml⁻¹. A 250-μl portion of this sample was loaded on a 20/25/30/35/40/50% w/w sucrose-density step gradient in buffer A and centrifuged for 20 h at 200,000g with a Beckman SW41 rotor. The fraction containing large membrane fragments was harvested from the 40/50% interface with a blunted hypodermic syringe and frozen at -20 °C, with 45% sucrose as cryoprotectant, until required for AFM.

Atomic force microscopy

The sample solution was adsorbed on the surface of freshly cleaved mica (Ted Pella, Redding, California, USA). A small drop of adsorption buffer (10 mM Tris-HCl pH 7.5, 150 mM KCl, 25 mM MgCl₂) was applied to the mica surface to ensure a firm attachment of the membranes, then 1 μl of the sample was injected into the thin film of adsorption buffer and left for 1–1.5 h. The sample was then gently washed with the recording buffer (10 mM Tris-HCl pH 7.5, 150 mM KCl) and placed on the AFM stage, where 100 μl of recording buffer was added to the liquid cell. A home-built AFM was used²⁸. Standard Si₃N₄ cantilevers (ThermoMicroscopes, Sunnyvale, California, USA) had a spring constant of 0.5 N m⁻¹ and operating frequencies of 25–35 kHz in liquid. AFM topographs were obtained with the use of tapping mode in liquid. The images with the highest resolution were achieved when the free tapping amplitude was 1–2 nm and the amplitude setpoint was adjusted to minimal forces, resulting in the damping of the free amplitude by only 5–10%. Images contained 256 × 256 pixels and were recorded at a typical line frequency of 5–7 Hz. Quantitative analysis of the AFM topographs and three-dimensional representation of the surface structures were performed with Scanning Probe Image Processor (Image Metrology ApS, Lyngby, Denmark).

Received 4 May; accepted 8 July 2004; doi:10.1038/nature02823.

- Deisenhofer, J., Epp, O., Miki, K., Huber, R. & Michel, H. Structure of the protein subunits in the photosynthetic reaction centre of *Rhodospseudomonas viridis* at 3 Å resolution. *Nature* **318**, 618–624 (1985).
- Allen, J. P., Feher, G., Yeates, T. O., Komiya, H. & Rees, D. C. Structure of the reaction center from *Rhodobacter sphaeroides* R26: The protein subunits. *Proc. Natl Acad. Sci. USA* **84**, 6162–6166 (1987).
- McDermott, G. *et al.* Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature* **374**, 517–521 (1995).
- Rozsak, A. W. *et al.* Crystal structure of the RC-LH1 core complex from *Rhodospseudomonas palustris*. *Science* **302**, 1969–1972 (2003).
- Walz, T., Jamieson, S. J., Bowers, C. M., Bullough, P. A. & Hunter, C. N. Projection structures of three photosynthetic complexes from *Rhodobacter sphaeroides*: LH2 at 6 Å, LH1 and RC-LH1 at 25 Å. *J. Mol. Biol.* **282**, 833–845 (1998).
- Jamieson, S. J. *et al.* Projection structure of the photosynthetic reaction centre-antenna complex of *Rhodospirillum rubrum* at 8.5 Å resolution. *EMBO J.* **21**, 3927–3935 (2002).
- Bahatyrava, S. *et al.* Flexibility and size heterogeneity of the LH1 light harvesting complex revealed by atomic force microscopy: functional significance for bacterial photosynthesis. *J. Biol. Chem.* **279**, 21327–21333 (2004).
- Fotiadis, D. *et al.* Structural analysis of the RC-LH1 photosynthetic core complex of *Rhodospirillum rubrum* using atomic force microscopy. *J. Biol. Chem.* **279**, 2063–2068 (2004).
- Siebert, C. A. *et al.* Molecular architecture of photosynthetic membranes in *Rhodobacter sphaeroides*: the role of PufX. *EMBO J.* **23**, 690–700 (2004).
- Vos, M. H., van Dorssen, R. J., Ames, J., van Grondelle, R. & Hunter, C. N. The organisation of the photosynthetic apparatus of *Rhodobacter sphaeroides*: studies of antenna mutants using singlet-singlet quenching. *Biochim. Biophys. Acta* **933**, 132–140 (1988).
- Aagaard, J. & Siström, W. R. Control of synthesis of reaction centre bacteriochlorophyll in photosynthetic bacteria. *Photochem. Photobiol.* **15**, 209–225 (1972).
- Monger, T. & Parson, W. W. Singlet-triplet fusion in *Rhodospseudomonas sphaeroides*. *Biochim. Biophys. Acta* **460**, 393–407 (1977).
- Hunter, C. N., Kramer, H. J. M. & van Grondelle, R. Linear dichroism and fluorescence emission of antenna complexes during photosynthetic unit assembly in *Rhodospseudomonas sphaeroides*. *Biochim. Biophys. Acta* **807**, 44–51 (1985).
- Visscher, K. J., Bergström, H., Sundström, V., Hunter, C. N. & van Grondelle, R. Temperature-dependence of energy-transfer from the long wavelength antenna BChl-896 to the reaction centre in *Rhodospirillum rubrum*, *Rhodobacter sphaeroides* (wt. and M21 mutant) from 77 to 177K, studied by

- picosecond absorption spectroscopy. *Photosyn. Res.* **22**, 211–217 (1989).
15. Beekman, L. M. *et al.* Trapping kinetics in mutants of the photosynthetic purple bacterium *Rhodobacter sphaeroides*: influence of the charge separation rate and consequences for the rate-limiting step in the light-harvesting process. *Biochemistry* **33**, 3143–3147 (1994).
 16. Frese, R. N. *et al.* The long-range supraorganization of the bacterial photosynthetic unit: A key role for PufX. *Proc. Natl Acad. Sci. USA* **97**, 5197–5202 (2000).
 17. Scheuring, S. *et al.* Nanodissection and high-resolution imaging of the *Rhodopseudomonas viridis* photosynthetic core complex in native membranes by AFM. *Proc. Natl Acad. Sci. USA* **100**, 1690–1693 (2003).
 18. Hess, S. *et al.* Temporally and spectrally resolved subpicosecond energy transfer within the peripheral antenna complex (LH2) and from LH2 to the core antenna complex in photosynthetic purple bacteria. *Proc. Natl Acad. Sci. USA* **92**, 12333–12337 (1995).
 19. Scheuring, S. *et al.* AFM characterization of tilt and intrinsic flexibility of *Rhodobacter sphaeroides* light harvesting complex 2 (LH2). *J. Mol. Biol.* **325**, 569–580 (2003).
 20. Katona, G., Andreasson, U., Landau, E. M., Andreasson, L. E. & Neutze, R. Lipidic cubic phase crystal structure of the photosynthetic reaction centre from *Rhodobacter sphaeroides* at 2.35 angstrom resolution. *J. Mol. Biol.* **331**, 681–692 (2003).
 21. Scheuring, S. *et al.* Structural role of PufX in the dimerization of the photosynthetic core-complex of *Rhodobacter sphaeroides*. *J. Biol. Chem.* **279**, 3620–3626 (2004).
 22. Sturgis, J. N. & Niederman, R. A. The effect of different levels of the B800–850 light-harvesting complex on intracytoplasmic membrane development in *Rhodobacter sphaeroides*. *Arch. Microbiol.* **165**, 235–242 (1996).
 23. Müller, D. J. *et al.* Observing membrane protein diffusion at subnanometer resolution. *J. Mol. Biol.* **327**, 925–930 (2003).
 24. Berry, E. A. *et al.* X-ray structure of *Rhodobacter capsulatus* cytochrome *bc_L*: Comparison with its mitochondrial and chloroplast counterparts. *Photosyn. Res.* (in the press).
 25. Feniouk, B. A., Cherepanov, D. A., Voskoboinikova, N. E., Mulkidjanian, A. Y. & Junge, W. Chromatophore vesicles of *Rhodobacter capsulatus* contain on average one F_0F_1 -ATP synthase each. *Biophys. J.* **82**, 1115–1122 (2002).
 26. Boekema, E. J., van Breemen, J. F., van Roon, H. & Dekker, J. P. Arrangement of photosystem II supercomplexes in crystalline macrodomains within the thylakoid membrane of green plant chloroplasts. *J. Mol. Biol.* **301**, 1123–1133 (2000).
 27. Niederman, R. A., Mallon, D. E. & Parks, L. C. Membranes of *Rhodopseudomonas sphaeroides*. VI. Isolation of a fraction enriched in newly synthesized bacteriochlorophyll α -protein complexes. *Biochim. Biophys. Acta* **555**, 210–220 (1979).
 28. Van der Werf, K. O. *et al.* Compact stand-alone atomic force microscope. *Rev. Sci. Instrum.* **64**, 2892–2897 (1993).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by grants from the BBSRC (UK) and the Netherlands Organisation for Scientific Research (NWO). This paper is dedicated to the memory of Prof. Dr. Bart de Grooth.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.N.H. (c.n.hunter@shef.ac.uk).

corrigendum

Germline stem cells and follicular renewal in the postnatal mammalian ovary

Joshua Johnson, Jacqueline Canning, Tomoko Kaneko, James K. Pru & Jonathan L. Tilly

Nature **428**, 145–150 (2004).

In this Article, we estimated that germline stem cells generate approximately 77 new primordial oocytes per day in ovaries of postnatal female mice. It has since been drawn to our attention by Ton Schumacher that a line of reasoning used in our study to verify this value is circular. However, the initial value for daily oocyte renewal obtained from mathematical modelling is not derived from a circular argument. Accordingly, this oversight does not alter any of the data or the conclusions in our paper. □

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

US Head Office, New York

345 Park Avenue South, 10th Floor,
New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager/ Corporations:

Peter Bless

Classified Sales Representatives:

Phone: +1 800 989 7718

Pennsylvania, New York,

Maryland/ Latin America/ NIH:

Michael McGillion

Midwest USA:

Wade Tucker

East USA/ Canada:

Janine Taormina

San Francisco Office

Classified Sales Representative:

Michaela Bjorkman

West USA/ West Corp. Canada

225 Bush Street, Suite 1453

San Francisco, CA 94104

Tel +1 415 781 3803

Fax +1 415 781 3805

e-mail: m.bjorkman@nature.sf.com

European Head Office, London

The Macmillan Building, 4 Crinan Street,
London N1 9XW, UK

Tel +44 (0) 20 7843 4961

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

e-mail: p.phelan@nature.com

o.wulffen@nature.com

Japan Head Office, Tokyo

MIG Ichigaya Building (5F),
19-1 Harakatemachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@nature.jp.com

naturejobs

Coping with the inevitable

It is a near-unavoidable paradox. Scientists pursue a career in science because they love research, but the more successful they become, the farther they find themselves from the bench. Instead of collating data, they get bogged down in the minutiae of management. What's worse, few find themselves prepared for such a role, even though running a lab, getting grants and cementing alliances all depend on management skills.

So where does one get such skills? Some people are naturals; most muddle through. But a few scientists who cut their scientific and management teeth in industry, and who have now returned to run academic institutions, are suggesting that some formal training may be in order.

Richard Sykes, former chairman of GlaxoSmithKline and now rector of Imperial College London, feels that there is a need for specialized programmes in the management of science, technology and engineering. Compared with an MBA, these courses would teach skills targeted at science and would create a smoother career path from student to postdoc to manager (see www.nature.com/naturejobs/channels/graduate/int_sir-richard-sykes.html). And Gail Naughton, dean of the College of Business Administration at San Diego State University, has launched a combined MBA-PhD programme that has internship, thesis and business-plan components (see *Nature* 430, 706-707; 2004).

Clearly, there is no one right way to teach scientific management, just as there is no one archetypal scientific manager. The different demands of small labs, large multidisciplinary efforts and multi-site initiatives all require different approaches. But putting management courses into the training mix of a PhD, then honing these skills during a postdoc, would help young scientists when they become principal investigators — allowing them to be ready for a role that their scientific skills will make inevitable.

Paul Smaglik
Naturejobs editor



Contents

CAREERS AND RECRUITMENT

Diabetes research
gains impetus p1064

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Driving back diabetes

As Western lifestyles spread around the world, diabetes has become an epidemic. Improved treatments are desperately needed, and the funding is there for those who may be able to help, says Ricki Lewis.

One of the earliest references to diabetes mellitus is in an Egyptian papyrus describing the sweet-smelling urine that is a symptom of the disease. Though spotted thousands of years ago, diabetes — in which the body either fails to produce enough insulin (type 1) or becomes unable to use it effectively (type 2) — was left untreated until 1920, when insulin was used to reverse the condition in dogs. Despite further advances, including recombinant human insulin and transplants of insulin-producing islet cells, new ways to treat this life-threatening condition are needed urgently. Its global incidence, according to the World Health Organization, has rocketed from 30 million in 1985 to more than 170 million and is likely to reach at least 300 million by 2025. This epidemic stems from a population that's both ageing and piling on weight.

"There will be a great increase in funding in both private and public sectors, due to ageing and to the explosive increase in obesity, which is a key factor in developing Type 2 diabetes," says George Vlasuk, vice-president, cardiovascular disease and metabolic disease, at Wyeth Research in Collegeville, Pennsylvania.

Some 90% of diabetes is type 2, or non-insulin dependent (NIDDM) — once known as 'adult-onset diabetes' but now appearing at all ages.



Ellen Feigal: increase will continue.

Fortunately, new approaches to untangling the causes of the disease are becoming available. Along with increased public and private funding, these are providing job opportunities for those who want to tackle diabetes, says Ellen Feigal, vice-president of clinical sciences and deputy scientific director at the non-profit Translational Genomics Research Institute (TGen) in Phoenix, Arizona. Mary Loeken, investigator at the Joslin Diabetes Center in Boston, agrees. "There will continue to be an increase in jobs in biomedical research in general, and diabetes in

particular," she says.

Diabetes research crosses the board and involves large drug companies, government laboratories, academic centres and non-profit research institutions. And jobs aren't restricted to the bench or clinic. TGen, for example, has openings for science writers, director of grants administration and laboratory managers.

CAREER SATISFACTION

Although diabetes can be understood at the molecular and cellular levels, symptoms occur at the organ level, and the condition is more common in certain groups. So a satisfying career in diabetes research can combine a specific expertise, such as in signal transduction, with attention to the bigger picture, such as epidemiology or population genetics.

The diversity of the 30 researchers in Vlasuk's group reflects that range. His team members have qualifications relevant to drug development, in biochemistry, cell biology, molecular biology, gene expression and animal pharmacological models.

At TGen, projects range from dissecting the genetic bases of diabetes and its complications, to pharmacogenetics, to nutrition. TGen is closely involved in the 'Healthy Avondale 2010' project, an effort to improve the health of a community. Some of TGen's projects include communities such as the Pima Indians, who have one of the highest rates of diabetes in the world.

Although much research is done in the clinic, there is a place for scientific expertise to bridge the gap between bench and bedside. Whereas clinical trials are largely the turf of physicians, a cell biologist might work in drug discovery or toxicology, says Deborah Anzalone, senior medical director working on a drug

DIABETES RESEARCH INSTITUTE

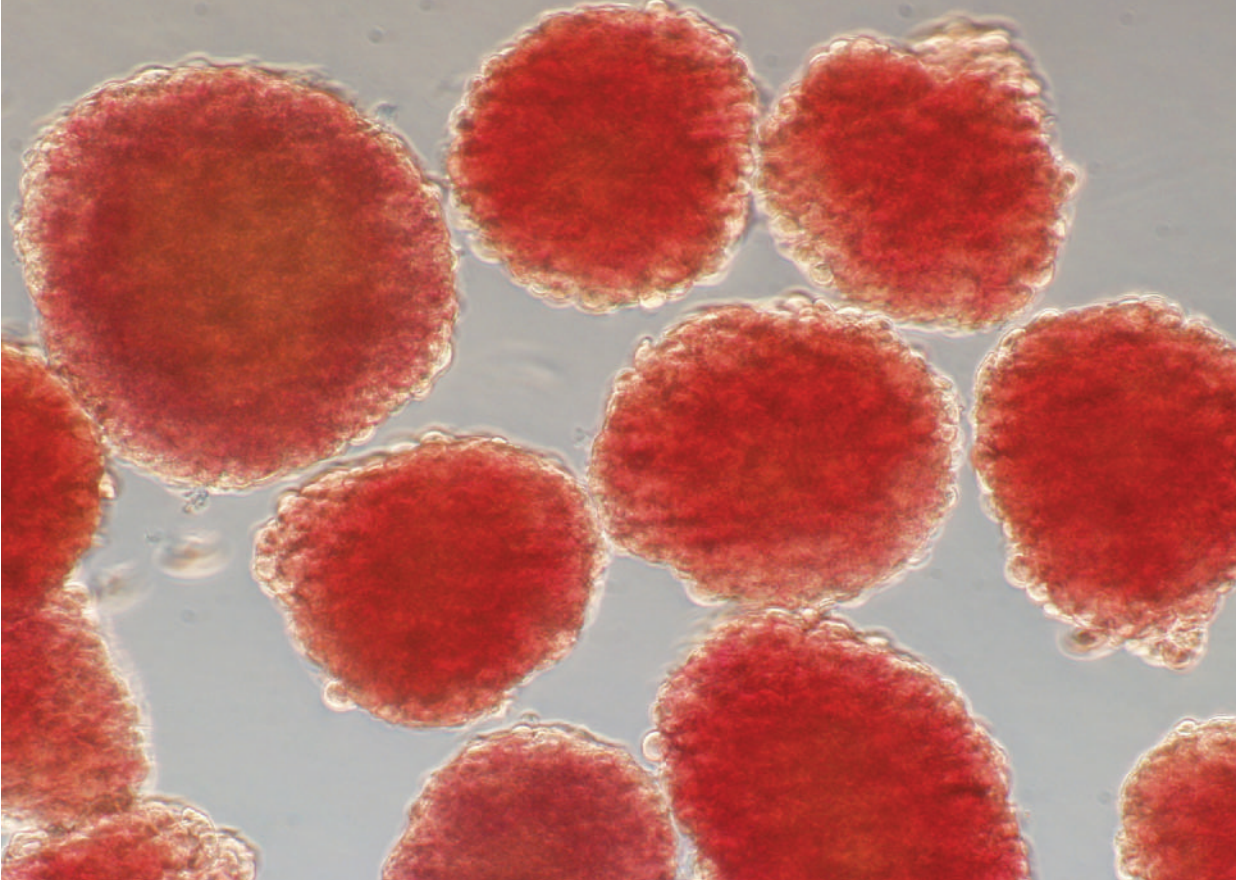
Wanted: islet specialists

With demand for islet transplants far exceeding supply, scientists are needed who can handle the delicate pancreatic insulin-producing tissue. The universities of Miami and Alberta offer training.

Camillo Ricordi, head of cellular transplantation at the Diabetes Research Institute at the University of Miami School of Medicine, works with several 'human-islet processing scientists'. The ideal candidate is a PhD level technologist with one to four years' experience in preparing this hard-to-sustain tissue. Technicians are also needed. The extensive training reflects the complexity of the material — the rare islets are dispersed and nestle intimately within the vast digestive portion of the pancreas. "An islet transplant is complex," says Ricordi, who in 1986 invented the device used worldwide to separate islets. "It involves more than a couple of technicians and a pancreas."

R. L.





DIABETES RESEARCH INSTITUTE

Islet of dreams: transplants of these insulin-producing cells could one day make insulin injections unnecessary, but success remains tantalizingly rare.

for type 2 diabetes at AstraZeneca in Boston. Medical schools increasingly stress scientific skills as preparation for careers in drug development. “A basic knowledge of biostatistics is fantastic to have, and biochemistry is also useful,” she adds.

Johanna Wolford, an associate investigator at TGen who collaborates with physicians and diabetes educators, says, “A paediatric endocrinologist might call me and say, ‘I’m seeing many more children with type 1 diabetes, what’s happening?’ An interface between clinical and laboratory research is critical.”

Large-scale studies also mix basic research and clinical investigation. The Finland–United States Investigation of NIDDM Genetics, for example, probes the genomes of Finns and Ashkenazi Jews for a particular gene variant associated with NIDDM in these groups. Investigators in the project have expertise in cell biology, human population genetics and new gene-scanning technologies, says Catherine McKeon, senior adviser for genetic research at the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in Bethesda, Maryland, which is one of the project’s funders.

The Type 1 Diabetes TrialNet investigates immune function and metabolism in close relatives of people with diabetes in 18 clinical centres in the United States, Canada, Europe and Australia. Many of the Immune Tolerance Network’s 80 physicians and clinical scientists at more than 40 institutions focus on islet transplants. The International Type 2 Diabetes Linkage Analysis Consortium probes genomes in several countries. For those interested in technology, the Endocrine Pancreas Consortium chronicles pancreas development using the PancChip DNA microarray.

Training positions are available at the NIDDK’s Phoenix Epidemiology and Clinical Research branch, for example. “We have positions for researchers wanting a year or two of research, perhaps between

university and a doctoral programme or medical school,” says the branch’s head, Clifton Bogardus.

With federal funding for human embryonic stem-cell research currently restricted in the United States, foundations are filling the void. The Juvenile Diabetes Research Foundation (JDRF) funds efforts in the United States, Australia, Canada, Israel, Sweden, Finland and Britain. Its 2004 \$93-million portfolio includes grants of \$100,000 to fund research too speculative to win money from the National Institutes of Health (NIH).

OBESITY RESEARCH EXPANDING

For 2005, the NIH is requesting \$150 million for a Special Type 1 Diabetes Initiative, and \$22 million for expanded research into obesity and diabetes. The multi-organ nature of diabetes is reflected in work on diabetic retinopathy at the US National Eye Institute in Bethesda, Maryland; on autoimmunity at the National Institute of Allergy and Infectious Diseases; and on diabetic neuropathy at the National Institute of Neurological Disorders and Stroke.

In Britain, the Medical Research Council supports diverse approaches to diabetes. The UK Prospective Diabetes Study, which ended in 1998, monitored health; the Protein Phosphorylation Unit takes a molecular approach. Genetic efforts map individual susceptibility genes and screen half-a-million people as part of the UK Biobank database project.

Now that islet-cell transplants work, and stem cells are poised to tackle the supply and rejection problems, diabetes may one day make history again — by becoming completely treatable. But that’s at least a decade away, say researchers. In the meantime, working to vanquish diabetes can provide a challenging and rewarding career. ■

Ricki Lewis is a freelance science writer in Scotia, New York.